Uncovering microsatellites' mutational mechanisms

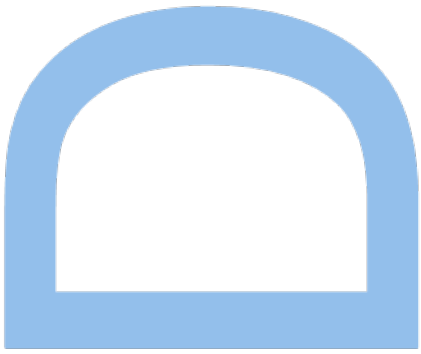
Ana Sofia Antão Sousa
Doutoramento em Biologia
Departamento de Biologia
2023

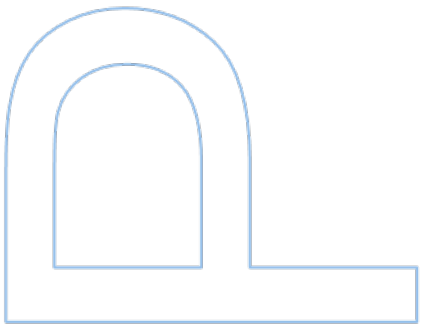
Orientador
Doutora Nádia Pinto
Instituto de Patologia e Imunologia Molecular da Universidade do Porto, IPATIMUP, Porto, Portugal
Instituto de Investigação e Inovação em Saúde, i3S, Universidade do Porto, Porto, Portugal
Centro de Matemática da Universidade do Porto, CMUP, Porto, Portugal

Coorientadores
Professora Doutora Leonor Gusmão
Laboratório de Diagnósticos por DNA, Universidade do Estado do Rio de Janeiro, LDD-UERJ, Rio de Janeiro, Brasil

Professor Doutor António Amorim
Instituto de Patologia e Imunologia Molecular da Universidade do Porto, IPATIMUP, Porto, Portugal
Instituto de Investigação e Inovação em Saúde, i3S, Porto, Portugal
Departamento de Biologia, Faculdade de Ciências da Universidade do Porto, FCUP, Porto, Portugal


U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO





Declaração de Honra

Eu, Ana Sofia Antão Sousa, inscrita no Programa Doutoral em Biologia da Faculdade de Ciências da Universidade do Porto declaro, nos termos do disposto na alínea a) do artigo 14.º do Código Ético de Conduta Académica da U.Porto, que o conteúdo da presente tese reflete as perspetivas, o trabalho de investigação e as minhas interpretações no momento da sua entrega.

Ao entregar esta tese, declaro, ainda, que a mesma é resultado do meu próprio trabalho de investigação e contém contributos que não foram utilizados previamente noutros trabalhos apresentados a esta ou outra instituição.

Mais declaro que todas as referências a outros autores respeitam escrupulosamente as regras da atribuição, encontrando-se devidamente citadas no corpo do texto e identificadas na secção de referências bibliográficas. Não são divulgados na presente tese quaisquer conteúdos cuja reprodução esteja vedada por direitos de autor.

Tenho consciência de que a prática de plágio e auto-plágio constitui um ilícito académico.

Ana Sofia Antão Sousa

Porto, 5 de julho de 2023

Index

Agradecimentos

Primeiramente gostaria de agradecer à Fundação para a Ciência e Tecnologia (FCT) por financiar o meu doutoramento através da bolsa de doutoramento SFRH/BD/136284/2018 concedida através do Concurso para Atribuição de Bolsas de Investigação para Doutoramento de 2018.

Um muito obrigada à minha orientadora Doutora Nádia Pinto por ter aceitado orientar-me uma vez mais. Obrigada por toda a disponibilidade e paciência que sempre tiveste, pelas oportunidades que fizeste questão de me proporcionar, e, acima de tudo, pela amizade.

Ao meu coorientador Professor Doutor António Amorim por me permitir desenvolver o meu trabalho no grupo do i3S que lidera, Genética Populacional e Evolução. Obrigada por ter sempre disponibilidade para rever todos os trabalhos e dar o seu grande contributo. Obrigada a todo o grupo de Genética Populacional e Evolução e a todos os colegas.

À minha coorientadora Professora Doutora Leonor Gusmão pela total disponibilidade que sempre demonstrou para esclarecer quaisquer dúvidas, pequenas ou grandes. Obrigada por me ter recebido no Laboratório de Diagnósticos por DNA no Rio de Janeiro e por me ter feito sentir em casa. Obrigada ao Professor Doutor Elizeu Carvalho e a todos os colegas do Laboratório de Diagnósticos por DNA no Rio de Janeiro por me terem recebido tão bem.

À Professora Doutora Vânia Pereira e à Faculty of Health and Medical Sciences, Department of Forensic Medicine em Copenhaga. Por me terem proporcionado a oportunidade de ganhar novos conhecimentos e experienciar uma nova cultura. À Sociedade Internacional de Genética Forense (ISFG) por me terem concedido a bolsa (Short Term Fellowship) que permitiu esta visita.

À Qiagen por me ter concedido o prémio Runner Up Young Investigator Award de 2019, que financiou a minha ida aos 28º e 29º congressos da Sociedade Internacional de Genética Forense (ISFG).

A todos os meus amigos e família, de coração! Por acreditarem em mim e estarem sempre do meu lado. Um obrigada muito especial à minha mãe Luísa, ao meu irmão Frederico e ao meu marido Samuel.

Obrigada!

Figure index

List of abbreviations

Bp: Base pair

CE: Capillary Electrophoresis

CODIS: Combined DNA Index System

DNA: Deoxyribonucleic Acid

FBI: Federal Bureau of Investigation

FSS: Forensic Science Service

GHEP-ISFG: Spanish and Portuguese Speaking Group of the International Society for Forensic Genetics

Indel: Insertion-deletion

MIA: Mendelian Incompatibilities Approach

Mis: Mendelian incompatibilities

MSY: Male-specific Region of the Y chromosome

PAR: Pseudoautosomal Region

PAR1: Pseudoautosomal Region 1

PAR2: Pseudoautosomal Region 2

PCR: Polymerase Chain Reaction

PD: Discrimination Power

PE: Exclusion Power

RFLP: Restriction Fragment Length Polymorphism

RNA: Ribonucleic Acid

SNP: Single Nucleotide Polymorphism

STR: Short Tandem Repeat

X-STR: X-chromosomal Short Tandem Repeat

Y-STR: Y-chromosomal Short Tandem Repeat

VNTR: Variable Number of Tandem Repeats

Resumo

Os *short tandem repeats* (STRs) são dos marcadores genéticos mais utilizados, sendo úteis em vários âmbitos, tais como antropologia, evolução e genética populacional e forense. As taxas de mutação associadas aos STRs são parâmetros necessários nos mais diversos contextos, sendo a precisão da sua estimativa crucial para, por exemplo, calcular o tempo de coalescência entre os alelos de um locus, ou análises de parentesco.

As mutações são eventos raros pelo que são necessárias quantidades substanciais de dados genéticos para obter estimativas precisas das respetivas taxas de ocorrência. Assim, um dos objetivos deste trabalho prendeu-se com a recolha de dados já publicados, bem como com a produção de novos relativos às taxas de mutação dos STRs dos cromossomas Y e X. Nomeadamente, foram obtidos novos dados correspondentes a 2.225 duos pai-filho e 2.273 trios pai-mãe-filha e recolhidos dados da literatura relativos a 25.729 duos pai-filho e 3.748 trios pai-mãe-filha para marcadores dos cromossomas Y e X, respetivamente. Estes dados permitiram concluir que a taxa de mutação por marcador é altamente variável (para o cromossoma Y entre 0,0005 para os marcadores DYS438 e DYS643, a 0,0170 para o marcador DYS547, e para o cromossoma X entre 0,0003 para o marcador DXS7133 a 0,0107 para o marcador DXS10135).

O tipo de mutação de alteração de comprimento mais comum em STRs é por um passo, ou seja, uma mutação que envolve o ganho ou perda de uma repetição, sendo mutações que envolvem mais de uma repetição menos frequentes. A taxa de mutação de STRs é influenciada por vários fatores, tais como a idade e o sexo dos indivíduos, o comprimento do alelo de origem, ou a sequência do motivo repetitivo do marcador. Neste trabalho, analisámos alguns destes fatores e quantificámos a sua correlação. A sequência do motivo repetitivo de um STR mostrou estar correlacionada com a razão entre o número de mutações por um ou vários passos, mesmo quando têm o mesmo comprimento. Especificamente, tetranucleótidos com diferentes sequências de motivo repetitivo (GAAA e GATA) mostraram proporções muito diferentes de mutações por um e vários passos.

As taxas de mutação dos STRs são geralmente estimadas através da abordagem por incompatibilidades Mendelianas (MIA) onde são analisados duos e trios familiares de uma geração – pai e/ou mãe-filho(a). A taxa de mutação é estimada diretamente calculando a razão entre o número de incompatibilidades Mendelianas observadas e o

número total de meioses. No entanto, a observação de compatibilidade entre os genótipos familiares não equivale à ausência de mutação. Utilizando plataformas comuns de genotipagem (eletroforese), apenas os STRs do cromossoma Y de estrutura simples, localizados na sua região específica (MSY, do inglês male-specific region of the Y chromosome), permitem a identificação inequívoca de que alelo parental resultou em que alelo filial. Para todos os outros sistemas genéticos, as mutações podem ser ocultadas ou classificadas erroneamente. Por conseguinte, a MIA aplicada aos modos de transmissão genética do cromossoma X e autossomas pode implicar diversos vieses. Um dos objetivos deste trabalho foi medir estes enviesamentos e propor correções considerando a presença de mutações ocultas.

Os nossos resultados, obtidos através de simulações computacionais, mostraram que, geralmente, os duos familiares escondem mais mutações do que os trios, impedindo o agrupamento de dados provenientes de diferentes tipos de configurações familiares.

A distribuição das frequências alélicas da população analisada, o tipo de configurações familiares consideradas e o tipo de mutação simulada influenciam fortemente a proporção de mutações ocultas, bem como a relação entre as mutações por um só passo e as que envolvem números mais elevados de repetições. Nesta tese, são propostos enquadramentos estatísticos e ferramentas informáticas para corrigir taxas de incompatibilidade obtidas através da MIA em configurações familiares de uma geração, considerando a presença de mutações ocultas.

Finalmente, importa referir que detetámos um elevado número de publicações com resultados reportados incorretamente (por exemplo, com diferentes tipos de configurações familiares agrupados) ou incapazes de serem reutilizados devido à publicação apenas de estatísticas sumárias, não estando disponíveis os dados em bruto. Assim, é essencial que a comunidade científica elabore recomendações para a padronização da publicação de dados de mutação, para permitir a sua replicação, reutilização e reanálise.

Palavras-chave: STRs, taxa de mutação de STRs, modelação da mutação, mecanismos de mutação

Abstract

Short tandem repeats (STRs) are one of the most used genetic markers. They are useful in an array of fields, such as anthropology, evolution and population and forensic genetics, and their mutation rates are a required parameter. The accuracy of its estimation is crucial, for example, to calculate the coalescence time between the alleles of a locus, or for kinship analyses.

Substantial amounts of genetic data are needed to obtain accurate estimates of mutation rates since mutations are rare events. Thus, one of the aims of this work was to gather published data as well as to produce new data regarding X- and Y-chromosomal STR mutation rates. Namely, we obtained new data corresponding to 2,225 father-son duos and 2,273 father-mother-daughter trios and collected data from the literature concerning 25,729 father-son duos and 3,748 father-mother-daughter trios. Altogether, they showed that per marker mutation rate varied widely between markers (from 0.0005 for markers DYS438 and DYS643, to 0.0170 for marker DYS547 for the Y chromosome, and from 0.0003 for marker DXS7133 to 0.0107 for marker DXS10135 for the X chromosome).

Several factors are known to influence the mutation rate, such as the age and sex of the individuals, the length of the allele of origin, or the sequence of the repetitive motif of the STR. In this work we analyzed some of these factors and quantified their influence. The sequence of the repetitive motif of the STR was shown to be correlated with the ratio of single/multistep mutations, beyond just its length, as tetranucleotides with different repetitive motif sequences (GAAA and GATA) showed widely different ratios of single- to multistep mutations.

STR mutation rates are usually estimated through the Mendelian incompatibilities approach (MIA) where one generation familial duos and trios – parent(s)-child – are analyzed and the ratio between the number of observed Mendelian incompatibilities and the total number of meiosis is computed. Nonetheless, the absence of incompatibility does not equate to the absence of mutation. Using the common genotyping platforms (CE electrophoresis), only simple-structure Y-STRs in the male-specific region of the Y chromosome (MSY) allow for the unambiguous definition of which paternal allele resulted in which filial one. For all other genetic systems, mutations can be concealed or misclassified. Therefore, the MIA applied to the X-chromosomal and autosomal modes of genetic transmission always entails biases. One of the goals of this work was to measure these biases and propose corrections to account for the presence of hidden mutations.

Our results, obtained through simulations, showed that, generally, familial duos conceal more mutations than trios, preventing the pooling of different types of family data.
The allele frequency distribution in the population analyzed, the type of familial data considered and the number of steps in the simulated mutation highly influenced the proportion of hidden mutations as well as the ratio between single step mutations and those involving higher numbers of repeats. In this thesis, frameworks and informatics tools are proposed to account for the presence of hidden mutations and correct incompatibility rates obtained through the MIA in one-generational familial configurations.
Finally, we detected a high number of reports in the literature that were either incorrect (e.g., by pooling different types of family data) or unusable because of publishing only summary statistics, the raw data being unavailable. Hence, it is essential that the forensic community elaborate recommendations for the standardization of mutation data publication, to allow its replicability, reuse, and reanalysis.

Keywords: STRs, STR mutation rates, mutation modeling, mutation mechanisms.

1. Introduction

1.1. Forensic Genetics

Forensic genetics is the application of genetics in the prevention, investigation, and resolution of legal matters. Namely, forensic genetics can be of use in the investigation phase, before the conflict between parties has entered the legal process, as in most cases of paternity tests; to assist in the preparation of the final process, as in the case of pleadings and dismissal of cases; or as a dissuading factor, as an individual is less likely to commit a crime if he/she believes he/she could be caught (Amorim & Budowle, 2017). A more comprehensive definition of forensic genetics may, in the future, include the fields of proteomics and epigenetics.

Karl Landsteiner gave the first step towards the establishment of the forensic genetics field in 1900 by describing the ABO blood grouping system (Karl Landsteiner, 1901). It was not until 1924, however, that Felix Bernstein proved that the system's transmission followed the rules of Mendelian inheritance, that it became evident that this information could be useful in paternity and crime cases (Bernstein, 1924). However, these serological tools were hindered by the amount of sample necessary to perform these analyses, and by the fact that proteins are prone to degradation upon exposure to the environment.

In the 1960s and 1970s, techniques such as those using restriction enzymes to detect Fragment Length Polymorphisms (RFLPs), Sanger sequencing (Sanger *et al.*, 1977), and Southern blotting (Southern, 1975), were developed and permitted scientists to examine DNA sequences. The first highly polymorphic locus was reported in the 1980s (Wyman & White, 1980).

In 1985, Alec Jeffreys discovered repetitive DNA sequences in the human genome. These sequences, which became known as Variable Number of Tandem Repeats (VNTR), vary among individuals, and allowed the performance of human identification tests (Jeffreys *et al.*, 1985).

Kary Mullis developed the method of amplification of DNA stretches, Polymerase Chain Reaction (PCR) in 1986 (Mullis *et al.*, 1986). PCR is an enzymatic methodology used to specifically amplify a single or few copies of a target DNA fragment into millions of copies. This technique promoted the development of DNA analyses on minute, vestigial samples.

The rapid progress in DNA analyses includes improvements in DNA extraction and quantification methodology and the development of commercial kits and equipment for the detection of DNA polymorphisms.

There are several contexts in which the comparison of genetic profiles and the correct quantification of the likelihood of their occurrence under various hypotheses is crucial, such as in:

- Kinship analyses, i.e., the analysis regarding a specific pedigree linking two or more individuals (Brustad *et al.*, 2021);
- Comparison of two genetic profiles in criminal investigation, e.g., a crime scene sample and the one from a suspect (Ralf & Kayser, 2021);
- Genetic genealogy and ancestry tests, through maternal or paternal lineage investigation (mtDNA and Y-chromosomal analyses, respectively) and Ancestry Informative Marker (AIM) analyses (Parson *et al.*, 2008; Romanini *et al.*, 2015; Toscanini *et al.*, 2016);
- Historical research, missing person, and disaster victim identification cases, through comparison of the genetic profile obtained from one sample and the genetic profile obtained from personal belongings or biological relatives (Bai *et al.*, 2022; Brenner & Weir, 2003; Gill *et al.*, 1994; Leclair *et al.*, 2001);
- Clinical diagnosis of genetic disorders (Chien *et al.*, 2022);
- Clinical investigation, as in the identification of cell lines to authenticate scientific research (Lv *et al.*, 2022).

While most of these analyses involve human DNA, non-human DNA may also be of use in forensic genetics (Arenas *et al.*, 2017). For example, in demonstrating the involvement or not of an individual suspected of a crime (Sensabaugh & Kaye, 1998). Domestic animals deposit hair where they live, and the hairs can be used to place a suspect at a scene. Likewise, proving a certain botanical sample came from a particular plant can help demonstrate the link of a suspect to a crime scene or help prove that a deceased body may have been moved. Moreover, stolen, or trafficked animals or illegally obtained meat (e.g., endangered species or poaching) can be identified through DNA testing (Giovambattista *et al.*, 2001).

Another example is the comparative genome sequencing for investigating infectious diseases, as was achieved for *Bacillus anthracis* (anthrax) with whole-genome sequencing (Read *et al.*, 2002, 2003). Additionally, phylogenetic analyses of viral strains of HIV have been admitted and used in court as evidence (Metzker *et al.*, 2002).

1.2. Genetic markers in forensic genetics

Genetic markers or polymorphisms are naturally occurring changes in the DNA sequence, where at least two alleles have frequencies greater than 1% in the population (Pereira & Gusmão, 2016). Several types of polymorphisms may be used in forensic genetics depending on the type of analyses intended. Genetic markers are present in coding and non-coding regions of the genome. However, for forensic identification purposes, markers from non-coding regions are chosen to avoid ethical concerns, as markers present in non-coding regions are less likely to convey information associated with diseases or genetic predispositions (Schneider, 1997).

Markers located in coding regions can, however, be useful in providing investigative leads in cases where no suspect was identified, namely, to determine phenotypic traits (forensic DNA phenotyping – FDP), such as appearance, ancestry, and age (see, for example, Kayser, 2015; Kayser & de Knijff, 2011; Phillips, 2015; Schneider *et al.*, 2019; Xavier *et al.*, 2022).

Genetic markers can be divided into recombining markers, for which meiotic recombination can occur – autosomal and X-chromosomal markers (the latter occurring solely in the female meiosis); and non-recombining markers, where no meiotic recombination normally occurs – Y-chromosomal markers (in the male-specific region, MSY, but occurring in the Pseudo Autosomal Regions, PARs) and mitochondrial markers. Consequently, the former category allows for individual identification while the latter ones only permit the identification of paternal or maternal lineages, respectively.

Further, genetic markers can be categorized into two main classes, bi-allelic, and multi-allelic markers. Bi-allelic markers are *loci* that present two possible alleles (A and B, for example), hence, three potential genotypes may exist in the population in diploid loci (AA, BB, and AB, as per the example). These include most Single Nucleotide Polymorphisms (SNPs) and Insertion/Deletion Polymorphisms (InDels). Multi-allelic markers possess several, more than two, polymorphic variants per *locus*, hence numerous genotypes do exist in the population. These include microsatellites or Short Tandem Repeats (STRs), the primary choice in the field of forensic genetics.

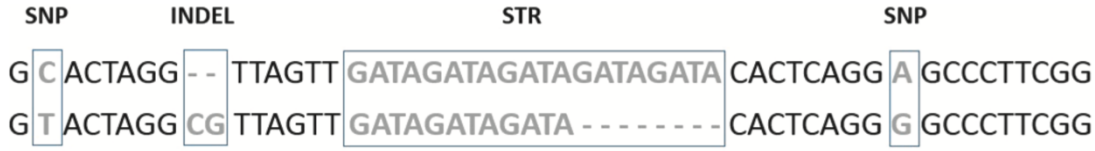
STRs consist of a tandemly arrayed 1-6 base pairs (bp) motif and span a median of 25 bp (see Figure 1). Approximately 700,000 STRs exist in the human genome, occupying 1% of its total length (Willems *et al.*, 2016). The average mutation rate of STRs is around 2.1×10^{-3} (Brinkmann *et al.*, 1998) making them very polymorphism genetic markers. STRs will be discussed further in section 1.2.1.

Single Nucleotide Polymorphisms (SNPs) are the simplest type of polymorphism. They arise when a point mutation i.e., a change involving a single nucleotide, occurs during meiosis when the cell undergoes DNA replication (see Figure 1). A mutation is characterized by a change in the DNA of an organism. If a mutation occurs in the germline (germline mutation) it can be passed down to offspring, otherwise, somatic mutations occur in non-germline body cells are not inheritable. Mutations will be discussed further in section 1.3. SNPs usually have only two alleles and therefore are not very polymorphic, being less used than the STRs for forensic purposes. However, since they are so abundant in the human genome, high power of discrimination can be obtained by typing enough SNPs. Moreover, they can be typed in shorter amplicons, which makes them useful in the analysis of degraded samples. SNPs are much more stable than STRs, with lower mutation rates (magnitude of about 10^{-8}, (Nachman & Crowell, 2000). To achieve the discrimination power of 10 STRs, it is estimated that 50 to 80 SNPs would be necessary (Amorim & Pereira, 2005).

Another form of bi-allelic polymorphism is an insertion-deletion or indel. An indel is the insertion or deletion of a DNA fragment from one to hundreds of nucleotides (Lupski *et al.*, 1996)(see Figure 1), however the largest proportion of indels are those with relatively few nucleotides in allele-length difference. James Weber and colleagues described over 2000 bi-allelic indels in the human genome, 71% of these possessing 2-, 3-, or 4 nucleotides of length differences with only 4% having a difference greater than 16 nucleotides (Weber *et al.*, 2002). Alike SNPs, indels are stable and plentiful markers, with lower mutation rates than that of STRs. These markers are easily typed and may be valuable for genetic analyses including human identity testing (Pereira *et al.*, 2009).

In conclusion, STRs are the method of choice when assessing genetic diversity in populations or in identification of individuals due to high levels of intra-population diversity. SNPs might be more adequate when lineage reconstruction is needed, or when analyzing degraded samples or samples with low amounts of DNA, since SNPs can be analyzed in short fragments. Indels are also typically used in samples with highly degraded DNA as the amplicon size is usually bellow 160bp. Indels can also be of use in kinship cases, with the particularity of having lower mutation rates that STRs.

Figure 1: Examples of different types of DNA markers: single nucleotide polymorphism (SNP), insertion/deletion (indel) and short tandem repeat (STR).



Source: (Pereira & Gusmão, 2016)

1.2.1. Short tandem repeats (STRs)

Short tandem repeats (STRs), also known as microsatellites, are tandemly arrayed DNA fragments of 1-6 base pairs per repeat motif. They account for 3% of the human genome (Lander *et al.*, 2001) and usually range from 50 to 300 base pairs (bp). Variation amongst alleles is generally caused by a difference in the number of repeat units, which results in alleles that vary in repeat length with a fixed number of base pairs. For example, tetranucleotide markers vary in length amongst themselves by multiples of 4 bp. STRs have been extensively used in diverse fields such as population and forensic genetics, anthropology, and evolution (see, for example, (Flores-Espinoza *et al.*, 2021; He *et al.*, 2017; Srithawong *et al.*, 2021).

STRs can be defined as simple, compound, or complex, depending on the structure of the sequence (Bar *et al.*, 1997). Simple STRs only show variation in the number of repeats without additional sequence variation. Compound STRs consist of several adjacent repeats of the same repeat unit length, and complex STRs contain repeats of variable length as well as sequences (see Figure 1) (Urquhart *et al.*, 1994). This structural complexity has implications for allele diversity and mutation rates of STRs as will be further explored in section 2.

Figure 2: Exemplification of simple, compound, and complex STRs.

Class	Locus	Example of allele sequence
Simple	CSF1PO	$[TCTA]_8$
Compound	vWA	$[TCTA][TCTG]_4[TCTA]_{13}$
Complex	D1S1656	$[TAGA]_{12}[TAGG]_1[TG]_5$

Tetranucleotide STRs are abundant, present less slippage than their di- or trinucleotide counterparts and, thus, are easier to type and more useful in mixture detection, where minor allelic contributions may be concealed by slippage artifacts. Thus, they have been preferred when designing commercial forensic kits (Pereira & Gusmão, 2016). Nonetheless, tri- or pentanucleotide STRs are also included in commercial kits.

Markers selected for forensic purposes are chosen based on the ability to discriminate between samples/individuals. Hence, the value or informative power of a marker (or set of markers) in identification is measured as the probability of two random (unrelated) individuals having different genetic profiles: the so-called Discrimination Power (PD), (or its complement, the Matching Probability) or, in kinship testing, through the probability of two random (unrelated) individuals being excluded as being related as parent-offspring: Exclusion Power (PE) (Pinto *et al.*, 2013). The latter is however a biased parameter in that it overlooks the positive evaluative value of the evidence and is prone to the misleading effect of silent alleles and mutation.

STRs fulfill all requirements for a forensic marker: their typing is robust, allowing for successful analyses in numerous biological materials; genotyping methods and nomenclature are well standardized so the results are easily compared between laboratories; they have a high power of discrimination, especially when analyzing several *loci* at once (PCR multiplexing); they are sensitive, enabling successful analyses even from trace, vestigial samples; the technology involved in the analyses is cheap, effortless and allows for automation; and a large number is located in non-coding regions of the genome, being less prone to deviations to the Hardy-Weinberg equilibrium caused by selective pressure.

The first multiplex PCR, described in 1988, was intended as a protocol to detect the dystrophin gene (Chamberlain *et al.*, 1988).

The first STRs used in forensic casework were selected by the UK Forensic Science Service (FSS), in 1994, for a quadruplex amplification system of four tetranucleotide markers: TH01, vWA, FES/FPS, and F13A1 (Kimpton *et al.*, 1994). These markers were selected based on their simple structure and the tendency to present several alleles differing by four base pairs. However, due to the low number of markers present in the kit, the power of discrimination was eventually deemed insufficient for the original purpose. So, in 1997, the USA Federal Bureau of Investigation (FBI) selected 13 autosomal loci to form a database of genetic profiles from federal, state, and local forensic laboratories, the Combined DNA Index System (CODIS). In 2017, other 7 STRs were added to the set. Most commercially available forensic kits include the core CODIS STR markers (Butler, 2006).

1.2.1.1. Autosomal STRs – applications in forensic genetics

The human genome is composed of 23 pairs of chromosomes. Twenty-two of these are autosomes, each pair containing one chromosome inherited from the mother and one from the father. In each generation, the alleles present in these chromosomes are shuffled, making it practically impossible that two random individuals to share the same genotype. On the other hand, any male or female individuals maternally or paternally related will share alleles by descent (shared alleles inherited from a common ancestor) with a probability that depends on the coefficient of co-ancestry between them (Pinto *et al.*, 2010, 2011). Commercial kits of autosomal markers present high power of discrimination, and their methods and protocols are well established, with guidelines on how to correctly report data (Bar *et al.*, 1997; Gill *et al.*, 2012; Gjertson *et al.*, 2007; Morling *et al.*, 2002; Prinz *et al.*, 2007). Since these markers allow for the identification of individuals, they are used to establish the identity of missing persons, analyze kinship hypotheses, or link persons of interest to crime scenes, with reliable answers for most cases. However, since they do not discriminate between individuals' sex, other complementary markers with sex influenced mode of transmission, may be required in special cases.

1.2.1.2. Heterosomal STRs

Complex cases may occur where autosomal genetic profiles are incomplete, or the interpretation of the results obtained is not straightforward which might require genetic markers with different characteristics or modes of transmission.

The human female and male genomes differ on the 23rd chromosomal pair, the sex-determining chromosomes, the so-called heterosomes. Females possess two X chromosomes while males possess one X chromosome and one Y chromosome. This particularity implies that the genetic transmission of the X chromosome in humans happens in a haplodiploid fashion. The transmission happens similarly to the autosomes in females, while in males there is no recombination between the X and Y chromosomes except at the pseudoautosomal regions (PAR1 and PAR2), two small X/Y homologous regions at the chromosomes' extremities.

The Y chromosome, exclusive to males, is transmitted in a haploid manner through the paternal lineage, from father to son, without recombination in most of its extension albeit

the PAR 1 and 2, which are responsible for the correct pairing of the X and Y chromosomes during recombination (Monteiro *et al.*, 2021). Hence, males from the same paternal lineage share the Y-chromosomal haplotype, unless mutations occur.
Combining the results obtained from markers with different types of genetic transmission has been successful when addressing the more complex cases (Coble *et al.*, 2009).

1.2.1.2.1. X-chromosomal STRs – applications in forensic genetics

The X chromosome accounts for around 5% of the human genome and holds upward of 1200 genes (Schwartz, 2013). One of the two female X chromosomes undergoes transcriptional inactivation to equalize the gene dosage between males and females. This phenomenon was first addressed by the Lyon hypothesis (Lyon, 1961), stating that only one of the female X chromosomes is active per cell. The chromosomal inactivation occurs in a randomly selected X chromosome when RNA transcribed from the *Xist* gene coats the chromosome from which it was expressed. The inactivation is stable, and the same X chromosome remains inactive in subsequent cell generations. However, this inactivation is incomplete, with some genes escaping the process (Lopes *et al.*, 2011). This mechanism explains why X aneuploidies (from X monosomy to heterokaryotypes XXY, XXXY) are compatible with viable individuals.
The X-chromosomal markers' particular mode of genetic transmission, where men possess only one X chromosome which they transmit to a daughter without recombination and women possess two X chromosomes which recombine before being transmitted to a child, implies that the value of these markers in forensic analyses varies depending on the sex of the individual, as well as the nature of the sample. Autosomal and X-chromosomal markers have the same power to discriminate females. However, since men only possess one X chromosome, the power to discriminate them is lower for X-chromosomal than for autosomal markers, considering equally diverse loci.
In mixtures with contributors of different sexes, X-chromosomal markers are particularly useful when there is a need to identify female traces in a male background (Szibor, 2007; Szibor *et al.*, 2003). For the male and female profiles to be equally discriminatory, the female would have to be homozygous for all markers analyzed, which although theoretically possible is highly unlikely due to the polymorphic nature of the STRs selected for forensic casework.
The main usage of X-chromosomal markers is in complex kinship cases, more specifically in incest cases where autosomal data might not resolve the kinship (Gomes *et al.*, 2012; Pinto *et al.*, 2011; Szibor, 2007; Szibor *et al.*, 2003) or in cases where the

putative fathers are father and son, and the child is a female. Since father and son do not share any X-chromosomal markers identical by descent and transmit the X-haplotype to the daughter without any change excluding mutations, the analyses of X-chromosomal markers will possibly help resolve the case. Likewise, in cases where the putative father is not available and a close relative has to be analyzed instead, X-chromosomal markers are particularly powerful in the cases of full- or paternal half-sisters or paternal grandmother-granddaughter cases, since these pairs of females share, unless mutation, the haplotype of one of their X chromosomes.

X-chromosomal markers can also be of relevance in the analyses of pedigrees that are indistinguishable when using unlinked autosomal markers but not when using X-chromosomal markers, as is the case of the avuncular, half-siblings and grandparent-grandchild kinships. As it is frequently relevant in forensic casework, the incestuous father-child kinship where the father is also the father, brother or son of the child's mother is also indistinguishable, when analyzing the mother-child duo (Pinto *et al.*, 2011).

1.2.1.2.2. Y-chromosomal STRs – applications in forensic genetics

The Y chromosome is the smallest chromosome of the human genome. It contains around 58 million base pairs (bp) (Morton, 1991) and represents 2% of the male genome. The Y chromosome is male-specific, being transmitted from father to son through the paternal lineage (see above the PAR exception). The 23rd pair of chromosomes is composed of two X chromosomes in females, and in males there is one X and one Y chromosome. This condition, which is exclusive to males, is named hemizygosity.

The Y chromosome is composed of two regions: the X-Y homology domain implicated in the meiotic pairing, the pseudoautosomal regions PAR1 and PAR2, and the male-specific region (MSY). The PAR1 and PAR2 are in the extremities of the chromosome and have 2.6Mb and 0.32Mb, respectively. The MSY lacks homology in the X chromosome and, thus, is not involved in the meiosis. So, except for the pseudoautosomal regions, the Y chromosome does not recombine in most of its extension, it is transmitted from father to son in a block, implying that variation, if any, is necessarily introduced by the occurrence of mutations (Jobling *et al.*, 2019).

Y-STRs are widely used in forensic genetics since most perpetrators in violent crimes and sexual assaults are male. Nonetheless, due to the lack of recombination, Y-chromosomal markers can only enlighten the history of male lineages. They can help disclose kinship when autosomal analyses are ambiguous. However, it is impossible of

distinguish males from the same paternal lineage based on Y-STR analyses (except in the case of mutations). Hence, if possible, autosomes should also be tested.

Another challenge that arises due to the lack of recombination is that the genetic data originating from Y-STRs can be statistically challenging as they must be treated as haplotypes rather than individual allele frequencies. The statistical evaluation of an evidence depends, hence, on the approach used to calculate match probabilities as most databases contain a high proportion of rare or unique haplotypes (because of the limited size of the haplotype database, most haplotypes might be seen just once) (Andersen *et al.*, 2013; Brenner, 2010, 2014; Caliebe *et al.*, 2015; Krawczak, 2001; Roewer *et al.*, 2000). The inclusion of rapidly mutating STRs in commercial kits presents the advantage of increasing the chance of discriminating between two paternal relatives. Due to the high diversity of these markers, they are also useful for excluding false persons of interest, which, however, will enhance the need for greater sample sizes and coverage to attain accurate statistical evidence estimation in non-exclusion cases (Pereira & Gusmão, 2016). Furthermore, an increase in mutation rates presents a drawback in kinship analyses, as there is an increased probability of two paternally related individuals presenting different Y-chromosomal haplotypes and in individual identification, as an individual may suffer a somatic mutation and present two haplotypes.

1.3. STR mutation

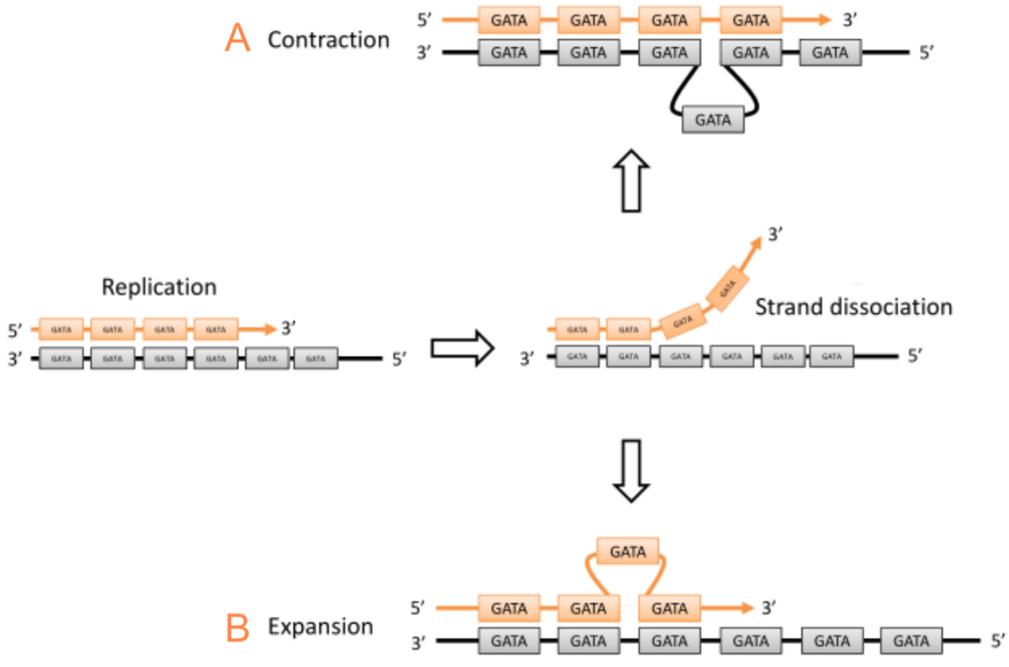
Mutations can be defined as a change in the genome, which can be transmitted to the offspring if occurring in the germinal line. These germinal mutations are likely to result in incongruences in the genotypes of parents and child, which is designated as a Mendelian incompatibility since it does not follow the transmission rules Gregor Mendel described in 1866 (Mendel, 1866).

It is generally accepted that random mutational processes create short tandem repeats, i.e., random point mutations followed by strand slippage which promotes the extension of the short primordial microsatellites (Levinson & Gutman, 1987; Schlotterer, 2000). Afterward, the main mechanism responsible for STR length variations, where STRs lose or gain repeats, is the polymerase template slippage (Schlötterer & Tautz, 1992; Strand *et al.*, 1993). During replication, the complex of the DNA polymerase and the newly synthesized DNA strand dissociate, temporarily, from the template DNA. As it is a repetitive sequence, the DNA polymerase may reassociate in a position one or several repeats distant from the original. If the polymerase slips forward, it skips a portion of the

template strand, resulting in the deletion or loss of a repeat (one or more) in the synthesized strand (see Figure 3, A), if it slips backward, it copies an extra repeat(s) into the new strand which results in the addition or gain of one (or more) repeat(s) (see Figure 3, B).

Single-step mutations, i.e., mutations involving the gain or loss of one repeat are thought to be much more probable than mutations involving several repeat changes, the so-called multistep mutations (Weber & Wong, 1993; X. Xu *et al.*, 2000). This premise is the foundation of the most commonly accepted mutation model, the single stepwise mutation model (or SMM), which will be discussed in further detail in section 1.4.3.

Figure 3: Schematic representation of the polymerase strand slippage.



Source: Adapted from (Hansson, 2018).

1.4. STR mutation rates estimation

1.4.1. Mendelian incompatibility approach (MIA)

STR mutation rates are usually estimated through the Mendelian incompatibilities approach (MIA) which consists of the genotyping of a large sample of one-generational familial pedigrees where the parentage was previously confirmed with a certain level of confidence and the corresponding weighing of Mendelian incompatibilities observed. In

the case of autosomes and the X chromosome, parent(s)-child trios or duos can be analyzed, and for the Y chromosome father-son duos. The standard practice considers that the relative frequency of the incompatibilities observed for a marker, considering the total number of meiosis, corresponds to its overall mutation rate. See, for example, (García *et al.*, 2017; Vieira *et al.*, 2017; Yang *et al.*, 2018).

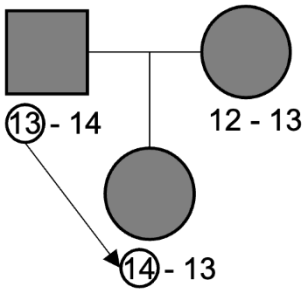
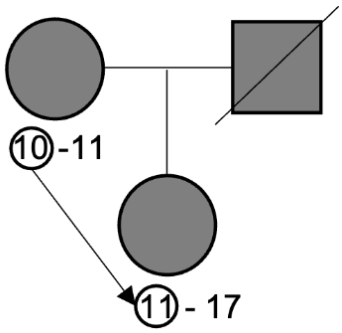
An alternate method to estimate Y-STR mutation is deep-rooting pedigrees. This approach, despite dispensing the need to run as many samples as father-son duos, has as disadvantages the difficulty to distinguish multistep mutations from several single-step mutations, and the occurrence of backward mutations (Claerhout *et al.*, 2018).

Several factors are known to influence mutation rates and dynamics. These are the allele length, repetitive motif size and sequence, parental age, and parental gender. Indeed, longer alleles mutate more frequently and tend to lose repeats, contrary to shorter alleles which mutate less frequently and tend to gain repeats (Wierdl *et al.*, 1997). Alleles with longer repetitive motifs tend to show lower mutation rates and, the repetitive motif sequence was also shown to influence mutation rates when comparing alleles with repetitive motifs of the same length (Ballantyne *et al.*, 2010; Brinkmann *et al.*, 1998). Lastly, mutations are more frequent in the paternal germline (compared to the maternal), a tendency that increases with the age of the father (Sun *et al.*, 2012).

1.4.2. STR mutation rate estimations: biases

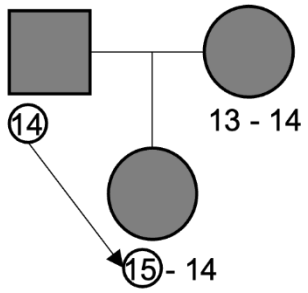
In the case of autosomal and X-chromosomal markers, the Mendelian incompatibilities approach methodology entails errors in the estimation of mutation rates. This is due to the assumptions regarding the number of steps involved in a mutation when counting observed incompatibilities. Some genotypical configurations may be explainable by a single-step mutation or no mutation at all. Since single-step mutations are presumed to be more frequent than multistep ones, the scenario which involves the least number of mutational steps is assumed, as no other, more complex, scenario is verifiable (Vicard & Dawid, 2004). Hence, except in the case of simple structured markers in haploid systems (the MSY), the MIA methodology involves biases resulting in the underestimation of mutation rates and multistep mutation rates, as mutations might not generate incompatibilities between parent(s) and child, causing 'hidden' or 'covert' mutations (Brenner, 2004; Chakraborty *et al.*, 1996; Slooten & Ricciardi, 2013; Vicard & Dawid, 2004) (see Figure 4).

Figure 4: Examples of hidden mutations occurring in a parents-child trio and a parent-child duo, at an autosomal STR. The arrows and circles indicate the allele transmissions involving a hidden mutation.

Familial genotypic configuration and mutation	**Mendelian compatibility interpretation (Absence of mutation)**
	Transmission of paternal allele 14 and maternal allele 13
	Transmission of maternal allele 11

Diploid and haplodiploid systems, i.e., autosomal, and X-chromosomal markers, respectively, imply yet another potential bias as they may not allow for the unequivocal assignment of the mutation to the correct parental origin (see Figure 5).

Figure 5: Example of ambiguous parental origin occurring in a parents-child trio, at an X-chromosomal STR. The arrows and circles indicate the allele transmissions involving a hidden mutation.

Familial genotypic configuration and mutation	**Mendelian interpretation (Ambiguous parental origin)**
	Transmission of mutated paternal allele 14 and maternal allele 14 or Transmission of mutated maternal allele 14 and paternal allele 14

These assumptions and inadequate assignments are exacerbating the problem. More mutations will be assumed as single-step or of paternal origin which will lead to an artificial increase in single-step and paternal mutation rates, which in turn will give more power to the assumptions and so on.

Hence, the likelihood with which a mutation results in Mendelian incompatibility depends not only on the parental genotypes but also on the type of familial configuration considered (Slooten & Ricciardi, 2013). For example, autosomal or chromosomal duos obligatorily hide more mutations than trios as there is less familial allelic information available in duos. Noteworthily, how if both the father and mother are homozygous for the same allele, any occurring mutation will be mandatorily visible.

1.4.3. STR mutation models

Several parameterized mutation models exist based on known incompatibility rates per marker and can be used to evaluate the link between two individuals in kinship investigations (Egeland *et al.*, 2016; Simonsson & Mostad, 2016). Mutation matrices, such as the one shown in Figure 6, can be built using parameters based on the different mutation models.

Figure 6: Example of a mutation matrix. Where n represents the number of possible alleles in the marker and m21 represents the probability that allele 2 is transmitted as mutated allele 1.

$$M=\begin{bmatrix} m_{11} & m_{12} & m_{13} & \cdots & m_{1n} \\ m_{21} & m_{22} & m_{23} & \cdots & m_{2n} \\ m_{31} & m_{32} & m_{33} & \cdots & m_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & m_{n3} & \cdots & m_{nn} \end{bmatrix}$$

Values across the diagonal are the probabilities of each allele being transmitted from father to son without mutation hence, the sum of these values should be close to 1. If R is the overall, per marker, mutation rate, assuming that the mutation probability is independent of the initial allele, then $\Sigma m_{ii} = 1 - R$.

1.4.3.1. The Equal Mutation Model

The Equal Mutation Model relies on the assumption that the probability of a mutation is the same regardless of the initial allele, and that the probability of said allele mutating to any other allele in the marker's allele range is the same (see Figure 7). While this model does not seem appropriate for STRs, as these assumptions are unrealistic given what is known of their mutational dynamics, it may be used due to its simplicity for computation, or when there is little information on which mutations are more probable for the marker in question, as in the case of SNPs.

Figure 7: Allele frequencies for marker TH01 for the population of Somalia (Kling *et al.*, 2014) and corresponding mutation matrix for the equal mutation model.

Allele	Frequency
6	0.182728
7	0.387985
8	0.178974
8.3	0.00125156
9	0.175845
9.3	0.0506884
10	0.0219024
12	0.000625782

	6	7	8	8.3	9	9.3	10	12
6	0.995	0.00071	0.00071	0.00071	0.00071	0.00071	0.00071	0.00071
7	0.00071	0.995	0.00071	0.00071	0.00071	0.00071	0.00071	0.00071
8	0.00071	0.00071	0.995	0.00071	0.00071	0.00071	0.00071	0.00071
8.3	0.00071	0.00071	0.00071	0.995	0.00071	0.00071	0.00071	0.00071
9	0.00071	0.00071	0.00071	0.00071	0.995	0.00071	0.00071	0.00071
9.3	0.00071	0.00071	0.00071	0.00071	0.00071	0.995	0.00071	0.00071
10	0.00071	0.00071	0.00071	0.00071	0.00071	0.00071	0.995	0.00071
12	0.00071	0.00071	0.00071	0.00071	0.00071	0.00071	0.00071	0.995

1.4.3.2. The Proportional to Frequency Mutation Model

The Proportional to Frequency Mutation model states that the probability of a mutation depends solely on the frequency of the mutated allele in the population and not on the initial allele. Specifically, when a mutation occurs, the mutated allele will result in a

random allele generated from the population frequency distribution (see Figure 8). This model is not biologically realistic (Vicard & Dawid, 2004).

Figure 8: Mutation matrix corresponding to the proportional to frequency mutation model obtained for the allele frequencies for marker TH01 for the population of Somalia (Kling *et al.*, 2014) presented in Figure 7.

	6	7	8	8.3	9	9.3	10	12
6	0.99455	0.00259	0.00119	0.00001	0.00117	0.00034	0.00015	0.00000
7	0.00122	0.99592	0.00119	0.00001	0.00117	0.00034	0.00015	0.00000
8	0.00122	0.00259	0.99453	0.00001	0.00117	0.00034	0.00015	0.00000
8.3	0.00122	0.00259	0.00119	0.99334	0.00117	0.00034	0.00015	0.00000
9	0.00122	0.00259	0.00119	0.00001	0.99451	0.00034	0.00015	0.00000
9.3	0.00122	0.00259	0.00119	0.00001	0.00117	0.99367	0.00015	0.00000
10	0.00122	0.00259	0.00119	0.00001	0.00117	0.00034	0.99348	0.00000
12	0.00122	0.00259	0.00119	0.00001	0.00117	0.00034	0.00015	0.99334

1.4.3.3. The Stepwise Mutation Model

The Stepwise Mutation Model considers the possibility of multistep mutations, despite considering single-step mutations the most common (Ohta & Kimura, 1973; Valdes *et al.*, 1993).

Mutations involving any pair of parental and filial alleles from the marker's allele range are considered possible, although the probability of mutation decreases with the increase of the difference in length of the (initial and final) alleles.

The Stepwise Mutation Model considers mutations between groups, of microvariant alleles (alleles with a non-integer number of repeats, for example, allele 14.3) and integer alleles (alleles with an integer number of repeats), as equally likely as mutations within groups (Figure 9). That is, a mutation from allele 13 to 14 is seen as equally likely as a mutation from allele 14 to 14.3, which is not biologically realistic.

Figure 9: Mutation matrix corresponding to the stepwise mutation model (unstationary) obtained for the allele frequencies for marker TH01 for the population of Somalia (Kling *et al.*, 2014) presented in Figure 7.

	6	7	8	8,3	9	9,3	10	12
6	0,995	0,0045	0,00045	4,50E-05	4,50E-06	4,50E-07	4,50E-08	4,50E-09
7	0,00236842	0,995	0,00236842	0,00023684	2,37E-05	2,37E-06	2,37E-07	2,37E-08
8	0,00022613	0,00226132	0,995	0,00226132	0,00022613	2,26E-05	2,26E-06	2,26E-07
8,3	2,25E-05	0,00022512	0,00225124	0,995	0,00225124	0,00022512	2,25E-05	2,25E-06

	6	7	8	8,3	9	9,3	10	12
9	2,25E-06	2,25E-05	0,00022512	0,00225124	0,995	0,00225124	0,00022512	2,25E-05
9,3	2,26E-07	2,26E-06	2,26E-05	0,00022613	0,00226132	0,995	0,00226132	0,00022613
10	2,37E-08	2,37E-07	2,37E-06	2,37E-05	0,00023684	0,00236842	0,995	0,00236842
12	4,50E-09	4,50E-08	4,50E-07	4,50E-06	4,50E-05	0,00045	0,0045	0,995

1.4.3.4. The Extended Stepwise Mutation Model

This model differs from the Stepwise Mutation Model as it considers the frequency for mutations involving a non-integer number of repeats (0.1, 0.2 or 0.3 bps in the case of tetranucleotide STRs) as far less likely than mutations involving an integer number of repeats (Figure 10). Hence, two mutation rates must be specified, the integer-length mutation rate (integer-length gain or loss) and the non-integer-length mutation rate (non-integer-length gain or loss).

Figure 10: Mutation matrix corresponding to the extended stepwise mutation model obtained for the allele frequencies for marker TH01 for the population of Somalia (Kling *et al.*, 2014) presented in Figure 7.

	6	7	8	8,3	9	9,3	10	12
6	0,99495	0,00450041	0,00045004	2,50E-05	4,50E-05	2,50E-05	4,50E-06	4,50E-08
7	0,00236956	0,99495	0,00236956	2,50E-05	0,00023696	2,50E-05	2,37E-05	2,37E-07
8	0,00022717	0,00227169	0,99495	2,50E-05	0,00227169	2,50E-05	0,00022717	2,27E-06
8,3	8,33E-06	8,33E-06	8,33E-06	0,99495	8,33E-06	0,005	8,33E-06	8,33E-06
9	2,36E-05	0,00023585	0,00235849	2,50E-05	0,99495	2,50E-05	0,00235849	2,36E-05
9,3	8,33E-06	8,33E-06	8,33E-06	0,005	8,33E-06	0,99495	8,33E-06	8,33E-06
10	4,13E-06	4,13E-05	0,00041288	2,50E-05	0,00412882	2,50E-05	0,99495	0,00041288
12	4,50E-07	4,50E-06	4,50E-05	2,50E-05	0,00045001	2,50E-05	0,00450005	0,99495

1.5. STR mutation rates: estimates correction

As previously referred, the estimation of mutation rates in systems other than the haploid (MSY) involves several biases. These biases are related to the imperfect way how mutations are counted, as only those resulting in Mendelian incompatibilities are considered. A most parsimonious reasoning is used, i.e., the scenario involving no

mutation and (if a Mendelian incompatibility is observed) the least number of mutational steps is the one considered to have occurred. As a result, mutation rates are underestimated, and single-step mutation rates are overestimated relatively to multistep ones. The first occurs due to the possibility of occurrence of hidden mutations which cannot be accounted for using this methodology without correction.

This issue was first addressed in 1996 by Chakraborty and colleagues (Chakraborty *et al.*, 1996) through computational simulation of a mutation in the paternal allele, considering familial trios. They found that STR mutation rates, without any correction, may be underestimated by as much as 60%.

In 2004, Vicard and Dawid (Vicard & Dawid, 2004), investigated the possible effects of biases as hidden mutations, differential mutation (paternal or maternal) and uncertain paternity, using family data. They concluded that correctly accounting for hidden mutations will increase the mutation rate estimates drastically.

Later, in 2008, Vicard and colleagues (Vicard *et al.*, 2008), advanced on this work with algebraic expressions and Bayesian networks to make inferences about mutation rates, considering several biases such as hidden mutations and uncertain paternity. They concluded that mutation models influence the estimation of mutation rates and suggest that the usage of data from different populations might be of interest to assess intrapopulation variation.

In 2013, Slooten and Ricciardi (Slooten & Ricciardi, 2013) used computational simulations to show that multistep mutations may have been greatly underestimated compared to single-step ones. Namely, in trios, mutations might be twice more likely than estimates obtained through the MIA methodology. Also, estimates were much more accurate when familial trios instead of duos were considered. The high variability of the biases' magnitude and type when different markers are analyzed was also shown. A mathematical correction was proposed for some autosomal STRs to account for the bias introduced by the presence of hidden mutations.

Introduction references

Amorim, A., & Budowle, B. (2017). Definition and Purpose. In *Handbook of Forensic Genetics: Biodiversity and Heredity in Civil and Criminal Investigation* (pp. 1–12). https://doi.org/10.1142/9781786340788_0001

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

Andersen, M. M., Eriksen, P. S., & Morling, N. (2013). The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *Journal of Theoretical Biology*, *329*, 39–51. https://doi.org/10.1016/J.JTBI.2013.03.009

Arenas, M., Pereira, F., Oliveira, M., Pinto, N., Lopes, A. M., Gomes, V., Carracedo, A., & Amorim, A. (2017). Forensic genetics and genomics: Much more than just a human affair. *PLoS Genetics*, *13*(9). https://doi.org/10.1371/JOURNAL.PGEN.1006960

Bai, Z., Zhang, N., Liu, J., Ding, H., Zhang, Y., Wang, T., Gao, J., & Ou, X. (2022). Identification of missing persons through kinship analysis by microhaplotype sequencing of single-source DNA and two-person DNA mixtures. *Forensic Science International: Genetics*, *58*, 102689. https://doi.org/10.1016/J.FSIGEN.2022.102689

Ballantyne, K. N., Goedbloed, M., Fang, R., Schaap, O., Lao, O., Wollstein, A., Choi, Y., van Duijn, K., Vermeulen, M., Brauer, S., Decorte, R., Poetsch, M., von Wurmb-Schwark, N., de Knijff, P., Labuda, D., Vézina, H., Knoblauch, H., Lessig, R., Roewer, L., … Kayser, M. (2010). Mutability of Y-Chromosomal Microsatellites: Rates, Characteristics, Molecular Bases, and Forensic Implications. *American Journal of Human Genetics*, *87*(3), 341. https://doi.org/10.1016/J.AJHG.2010.08.006

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 Society for Forensic Haemogenetics. *International Journal of Legal Medicine*, *110*(4), 175–176. https://doi.org/10.1007/S004140050061

Bernstein, F. (1924). Ergebnisse einer biostatistischen zusam-menfassenden betrachtung über die erblichen blutstrukturen des menschen. *Klinische Wochenschrift*, *3*(33), 1495–1497.

Brenner, C. H. (2004). Multiple mutations, covert mutations and false exclusions in paternity casework. *International Congress Series*, *1261*(C), 112–114. https://doi.org/10.1016/S0531-5131(03)01843-0

Brenner, C. H. (2010). Fundamental problem of forensic mathematics--the evidential value of a rare haplotype. *Forensic Science International. Genetics*, *4*(5), 281–291. https://doi.org/10.1016/J.FSIGEN.2009.10.013

Brenner, C. H. (2014). Understanding Y haplotype matching probability. *Forensic Science International. Genetics*, *8*(1), 233–243. https://doi.org/10.1016/J.FSIGEN.2013.10.007

Brenner, C. H., & Weir, B. S. (2003). Issues and strategies in the DNA identification of World Trade Center victims. *Theoretical Population Biology*, *63*(3), 173–178. https://doi.org/10.1016/S0040-5809(03)00008-X

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., & Rolf, B. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *American Journal of Human Genetics*, *62*(6), 1408–1415. https://doi.org/10.1086/301869

Brustad, H. K., Colucci, M., Jobling, M. A., Sheehan, N. A., & Egeland, T. (2021). Strategies for pairwise searches in forensic kinship analysis. *Forensic Science International: Genetics*, *54*, 102562. https://doi.org/10.1016/J.FSIGEN.2021.102562

Butler, J. M. (2006). Genetics and genomics of core short tandem repeat loci used in human identity testing. *Journal of Forensic Sciences*, *51*(2), 253–265. https://doi.org/10.1111/J.1556-4029.2006.00046.X

Caliebe, A., Jochens, A., Willuweit, S., Roewer, L., & Krawczak, M. (2015). No shortcut solution to the problem of Y-STR match probability calculation. *Forensic Science International. Genetics*, *15*, 69–75. https://doi.org/10.1016/J.FSIGEN.2014.10.016

Chakraborty, R., Stivers, D. N., & Zhong, Y. (1996). Estimation of mutation rates from parentage exclusion data: applications to STR and VNTR loci. *Mutation Research*, *354*(1), 41–48. https://doi.org/10.1016/0027-5107(96)00014-0

Chamberlain, J. S., Gibbs, R. A., Rainer, J. E., Nguyen, P. N., & Thomas, C. (1988). Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. *Nucleic acids research*, *16*(23), 11141-11156.

Chien, S. C., Chen, C. P., & Liou, J. der. (2022). Prenatal diagnosis and genetic counseling of uniparental disomy. *Taiwanese Journal of Obstetrics & Gynecology*, *61*(2), 210–215. https://doi.org/10.1016/J.TJOG.2022.02.006

Claerhout, S., Vandenbosch, M., Nivelle, K., Gruyters, L., Peeters, A., Larmuseau, M. H. D., & Decorte, R. (2018). Determining Y-STR mutation rates in deep-routing genealogies: Identification of haplogroup differences. *Forensic Science International. Genetics*, *34*, 1–10. https://doi.org/10.1016/J.FSIGEN.2018.01.005

Coble, M. D., Loreille, O. M., Wadhams, M. J., Edson, S. M., Maynard, K., Meyer, C. E., Niederstätter, H., Berger, C., Berger, B., Falsetti, A. B., Gill, P., Parson, W., & Finelli, L. N. (2009). Mystery Solved: The Identification of the Two Missing Romanov Children Using DNA Analysis. *PLOS ONE*, *4*(3), e4838. https://doi.org/10.1371/JOURNAL.PONE.0004838

Egeland, T., Kling, D., & Mostad, P. (2016). Relationship Inference with Familias and R: Statistical Methods in Forensic Genetics. *Relationship Inference with Familias and R: Statistical Methods in Forensic Genetics*, 1–243. https://doi.org/10.1016/C2014-0-01828-X

Flores-Espinoza, R., Paz-Cruz, E., Ruiz-Pozo, V. A., Lopez-Carrera, M., Cabrera-Andrade, A., Gusmão, L., & Burgos, G. (2021). Investigating genetic diversity in admixed populations from Ecuador. *American Journal of Physical Anthropology*, *176*(1), 109–119. https://doi.org/10.1002/AJPA.24341

García, M. G., Gusmão, L., Catanesi, C. I., Penacino, G. A., & Pinto, N. (2017). Mutation rate of 12 X-STRs from investigator Argus X-12 kit in Argentine population. *Forensic Science International: Genetics Supplement Series*, *6*, e562–e564. https://doi.org/10.1016/J.FSIGSS.2017.09.219

Gill, P., Gusmão, L., Haned, H., Mayr, W. R., Morling, N., Parson, W., Prieto, L., Prinz, M., Schneider, H., Schneider, P. M., & Weir, B. S. (2012). DNA commission of the International Society of Forensic Genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods. *Forensic Science International. Genetics*, *6*(6), 679–688. https://doi.org/10.1016/J.FSIGEN.2012.06.002

Gill, P., Ivanov, P. L., Kimpton, C., Piercy, R., Benson, N., Tully, G., Evett, I., Hagelberg, E., & Sullivan, K. (1994). Identification of the remains of the Romanov family by DNA analysis. *Nature Genetics*, *6*(2), 130–135. https://doi.org/10.1038/NG0294-130

Giovambattista, G., Ripoli, M. v., Lirón, J. P., Villegas Castagnasso, E. E., Peral-García, P., & Lojo, M. M. (2001). DNA Typing in a Cattle Stealing Case. *Journal of Forensic Sciences*, *46*(6), 1484–1486. https://doi.org/10.1520/JFS15176J

Gjertson, D. W., Brenner, C. H., Baur, M. P., Carracedo, A., Guidet, F., Luque, J. A., Lessig, R., Mayr, W. R., Pascali, V. L., Prinz, M., Schneider, P. M., & Morling, N. (2007). ISFG:

Recommendations on biostatistics in paternity testing. *Forensic Science International. Genetics*, *1*(3–4), 223–231. https://doi.org/10.1016/J.FSIGEN.2007.06.006

Gomes, C., Magalhaes, M., Alves, C., Amorim, A., Pinto, N., & Gusmao, L. (2012). Comparative evaluation of alternative batteries of genetic markers to complement autosomal STRs in kinship investigations: autosomal indels vs. X-chromosome STRs. *International Journal of Legal Medicine*, *126*(6), 917–921. https://doi.org/10.1007/S00414-012-0768-5

Hansson, O. (2018). *Development of computer software to characterise and simulate molecular biology processes used in forensic DNA profiling assays*.

He, G., Wang, Z., Wang, M., & Hou, Y. (2017). Genetic Diversity and Phylogenetic Differentiation of Southwestern Chinese Han: a comprehensive and comparative analysis on 21 non-CODIS STRs. *Scientific Reports 2017 7:1*, *7*(1), 1–8. https://doi.org/10.1038/s41598-017-13190-w

Jeffreys, A. J., Wilson, V., & Thein, S. L. (1985). Individual-specific 'fingerprints' of human DNA. *Nature 1985 316:6023*, *316*(6023), 76–79. https://doi.org/10.1038/316076a0

Jobling, M. A., Hurles, M., & Tyler-Smith, C. (2019). Human Evolutionary Genetics : Origins, Peoples and Disease. In *Human Evolutionary Genetics*. Garland Science. https://doi.org/10.1201/9780203487211

Landsteiner, K. (1901). On Agglutination Phenomena of Normal Human Blood. *Wiener Klinische Wochenschrift*, 1132–1134.

Kayser, M. (2015). Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. *Forensic Science International. Genetics*, *18*, 33–48. https://doi.org/10.1016/J.FSIGEN.2015.02.003

Kayser, M., & de Knijff, P. (2011). Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics 2011 12:3*, *12*(3), 179–192. https://doi.org/10.1038/nrg2952

Kimpton, C., Fisher, D., Watson, S., Adams, M., Urquhart, A., Lygo, J., & Gill, P. (1994). Evaluation of an automated DNA profiling system employing multiplex amplification of four tetrameric STR loci. *International Journal of Legal Medicine*, *106*(6), 302–311. https://doi.org/10.1007/BF01224776

Kling, D., Tillmar, A. O., & Egeland, T. (2014). Familias 3 – Extensions and new functionality. *Forensic Science International: Genetics*, *13*, 121–127. https://doi.org/10.1016/j.fsigen.2014.07.004

Krawczak, M. (2001). Forensic evaluation of Y-STR haplotype matches: A comment. *Forensic Science International*, *118*(2–3), 114–115. https://doi.org/10.1016/S0379-0738(00)00479-5

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Leclair, B., Frégeau, C. J., Bowen, K. L., & Fourney, R. M. (2001). Enhanced Kinship Analysis and STR-based DNA Typing for Human Identification in Mass Fatality Incidents: The Swissair Flight 111 Disaster. *Journal of Forensic Sciences*, *49*(5), JFS2003311-15. https://doi.org/10.1520/JFS2003311

Levinson, G., & Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution*, *4*(3), 203–221. https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A040442

Lopes, A. M., Arnold-Croop, S. E., Amorim, A., & Carrel, L. (2011). Clustered transcripts that escape X inactivation at mouse XqD. *Mammalian Genome : Official Journal of the International Mammalian Genome Society*, *22*(9–10), 572–582. https://doi.org/10.1007/S00335-011-9350-6

Lupski, J. R., Roth, J. R., & Weinstock, G. M. (1996). Chromosomal duplications in bacteria, fruit flies, and humans. *American Journal of Human Genetics*, *58*(1), 21. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1914933/

Lv, L., Luo, H., Yu, L., Tong, W., Jiang, Y., Li, G., Tong, G., Li, Y., & Liu, C. (2022). Identification and Characterization of Cell Lines HepG2, Hep3B217 and SNU387 as Models for Porcine Epidemic Diarrhea Coronavirus Infection. *Viruses*, *14*(12). https://doi.org/10.3390/V14122754

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

Mendel, G. (1866). Experiments in plant hybridization. *Verhandlungen Des Naturforschenden Vereins Brünn*. http://www.netspace.org./MendelWeb/

Metzker, M. L., Mindell, D. P., Liu, X. M., Ptak, R. G., Gibbs, R. A., & Hillis, D. M. (2002). Molecular evidence of HIV-1 transmission in a criminal case. *Proceedings of the National*

Academy of Sciences of the United States of America, *99*(22), 14292–14297. https://doi.org/10.1073/PNAS.222522599/SUPPL_FILE/5225FIG4.PDF

Monteiro, B., Arenas, M., Prata, M. J., & Amorim, A. (2021). Evolutionary dynamics of the human pseudoautosomal regions. *PLOS Genetics*, *17*(4), e1009532. https://doi.org/10.1371/JOURNAL.PGEN.1009532

Morling, N., Allen, R. W., Carracedo, A., Geada, H., Guidet, F., Hallenberg, C., Martin, W., Mayr, W. R., Olaisen, B., Pascali, V. L., & Schneider, P. M. (2002). Paternity Testing Commission of the International Society of Forensic Genetics: Recommendations on genetic investigations in paternity cases. *Forensic Science International*, *129*(3), 148–157. https://doi.org/10.1016/S0379-0738(02)00289-X

Morton, N. E. (1991). Parameters of the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(17), 7474. https://doi.org/10.1073/PNAS.88.17.7474

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, *51 Pt 1*(1), 263–273. https://doi.org/10.1101/SQB.1986.051.01.032

Nachman, M. W., & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, *156*(1), 297–304. https://doi.org/10.1093/GENETICS/156.1.297

Ohta, T., & Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research*, *22*(2), 201–204. https://doi.org/10.1017/S0016672300012994

Parson, W., Fendt, L., Ballard, D., Børsting, C., Brinkmann, B., Carracedo, Á., Carvalho, M., Coble, M. D., Real, F. C., Desmyter, S., Dupuy, B. M., Harrison, C., Hohoff, C., Just, R., Krämer, T., Morling, N., Salas, A., Schmitter, H., Schneider, P. M., … Brandstätter, A. (2008). Identification of West Eurasian mitochondrial haplogroups by mtDNA SNP screening: results of the 2006-2007 EDNAP collaborative exercise. *Forensic Science International. Genetics*, *2*(1), 61–68. https://doi.org/10.1016/J.FSIGEN.2007.08.007

Pereira, R., Phillips, C., Alves, C., Amorim, A., Carracedo, Á., & Gusmao, L. (2009). A new multiplex for human identification using insertion/deletion polymorphisms. *Electrophoresis*, *30*(21), 3682–3690. https://doi.org/10.1002/ELPS.200900274

Pereira, V., & Gusmão, L. (2016). Types of Genomes, Sequences and Genetic Markers (Repeats, SNPs, Indels, Haplotypes). In *Handbook of Forensic Genetics: Biodiversity and*

Heredity in Civil and Criminal Investigation (pp. 163–191). https://doi.org/10.1142/9781786340788_0009

Phillips, C. (2015). Forensic genetic analysis of bio-geographical ancestry. *Forensic Science International. Genetics*, *18*, 49–65. https://doi.org/10.1016/J.FSIGEN.2015.05.012

Pinto, N., Gusmão, L., & Amorim, A. (2011). X-chromosome markers in kinship testing: a generalisation of the IBD approach identifying situations where their contribution is crucial. *Forensic Science International. Genetics*, *5*(1), 27–32. https://doi.org/10.1016/J.FSIGEN.2010.01.011

Pinto, N., Gusmão, L., Egeland, T., & Amorim, A. (2013). Paternity exclusion power: Comparative behaviour of autosomal and X-chromosomal markers in standard and deficient cases with inbreeding. *Forensic Science International: Genetics*, *7*(2), 290–295. https://doi.org/10.1016/J.FSIGEN.2012.12.002

Pinto, N., Silva, P. v., & Amorim, A. (2010). General derivation of the sets of pedigrees with the same kinship coefficients. *Human Heredity*, *70*(3), 194–204. https://doi.org/10.1159/000316390

Prinz, M., Carracedo, A., Mayr, W. R., Morling, N., Parsons, T. J., Sajantila, A., Scheithauer, R., Schmitter, H., & Schneider, P. M. (2007). DNA Commission of the International Society for Forensic Genetics (ISFG): Recommendations regarding the role of forensic genetics for disaster victim identification (DVI). *Forensic Science International: Genetics*, *1*(1), 3–12. https://doi.org/10.1016/J.FSIGEN.2006.10.003

Ralf, A., & Kayser, M. (2021). Investigative DNA analysis of two-person mixed crime scene trace in a murder case. *Forensic Science International: Genetics*, *54*, 102557. https://doi.org/10.1016/J.FSIGEN.2021.102557

Read, T. D., Peterson, S. N., Tourasse, N., Baillie, L. W., Paulsen, I. T., Nelson, K. E., Tettelin, H., Fouts, D. E., Eisen, J. A., Gill, S. R., Holtzapple, E. K., Økstad, O. A., Helgason, E., Rilstone, J., Wu, M., Kolonay, J. F., Beanan, M. J., Dodson, R. J., Brinkac, L. M., … Fraser, C. M. (2003). The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria. *Nature*, *423*(6935), 81–86. https://doi.org/10.1038/NATURE01586

Read, T. D., Salzberg, S. L., Pop, M., Shumway, M., Umayam, L., Jiang, L., Holtzapple, E., Busch, J. D., Smith, K. L., Schupp, J. M., Solomon, D., Keim, P., & Fraser, C. M. (2002). Comparative genome sequencing for discovery of novel polymorphisms in Bacillus anthracis. *Science (New York, N.Y.)*, *296*(5575), 2028–2033. https://doi.org/10.1126/SCIENCE.1071837

Roewer, L., Kayser, M., de Knijff, P., Anslinger, K., Betz, A., Caglià, A., Corach, D., Füredi, S., Henke, L., Hidding, M., Kärgel, H. J., Lessig, R., Nagy, M., Pascali, V. L., Parson, W., Rolf, B., Schmitt, C., Szibor, R., Teifel-Greding, J., & Krawczak, M. (2000). A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forensic Science International*, *114*(1), 31–43. https://doi.org/10.1016/S0379-0738(00)00287-5

Romanini, C., Romero, M., Salado Puerto, M., Catelli, L., Phillips, C., Pereira, R., Gusmão, L., & Vullo, C. (2015). Ancestry informative markers: Inference of ancestry in aged bone samples using an autosomal AIM-Indel multiplex. *Forensic Science International: Genetics*, *16*, 58–63. https://doi.org/10.1016/J.FSIGEN.2014.11.025

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467. https://doi.org/10.1073/PNAS.74.12.5463

Schlotterer, C. (2000). Evolutionary dynamics of microsatellite DNA. *Chromosoma*, *109*(6), 365–371. https://doi.org/10.1007/S004120000089

Schlötterer, C., & Tautz, D. (1992). Slippage synthesis of simple sequence DNA. *Nucleic Acids Research*, *20*(2), 211. https://doi.org/10.1093/NAR/20.2.211

Schneider, P. M. (1997). Basic issues in forensic DNA typing. *Forensic Science International*, *88*(1), 17–22. https://doi.org/10.1016/S0379-0738(97)00079-0

Schneider, P. M., Prainsack, B., & Kayser, M. (2019). The Use of Forensic DNA Phenotyping in Predicting Appearance and Biogeographic Ancestry. *Deutsches Ärzteblatt International*, *116*(51–52), 873. https://doi.org/10.3238/ARZTEBL.2019.0873

Schwartz, C. E. (2013). X Chromosome. *Brenner's Encyclopedia of Genetics: Second Edition*, 352–354. https://doi.org/10.1016/B978-0-12-374984-0.01650-8

Sensabaugh, G., & Kaye, D. H. (1998). Non-human DNA evidence. *Jurimetrics J.*, *38*, 1–16.

Simonsson, I., & Mostad, P. (2016). Stationary mutation models. *Forensic Science International. Genetics*, *23*, 217–225. https://doi.org/10.1016/J.FSIGEN.2016.04.005

Slooten, K., & Ricciardi, F. (2013). Estimation of mutation probabilities for autosomal STR markers. *Forensic Science International. Genetics*, *7*(3), 337–344. https://doi.org/10.1016/J.FSIGEN.2013.01.006

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology*, *98*(3). https://doi.org/10.1016/S0022-2836(75)80083-0

Srithawong, S., Muisuk, K., Srikummool, M., Kampuansai, J., Pittayaporn, P., Ruangchai, S., Liu, D., & Kutanan, W. (2021). Close genetic relationship between central Thai and Mon people in Thailand revealed by autosomal microsatellites. *International Journal of Legal Medicine*, *135*(2), 445–448. https://doi.org/10.1007/S00414-020-02290-4

Strand, M., Prolla, T. A., Liskay, R. M., & Petes, T. D. (1993). Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature 1993 365:6443*, *365*(6443), 274–276. https://doi.org/10.1038/365274a0

Sun, J. X., Helgason, A., Masson, G., Ebenesersdóttir, S. S., Li, H., Mallick, S., Gnerre, S., Patterson, N., Kong, A., Reich, D., & Stefansson, K. (2012). A direct characterization of human mutation based on microsatellites. *Nature Genetics*, *44*(10), 1161–1165. https://doi.org/10.1038/NG.2398

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

Szibor, R., Krawczak, M., Hering, S., Edelmann, J., Kuhlisch, E., & Krause, D. (2003). Use of X-linked markers for forensic purposes. *International Journal of Legal Medicine*, *117*(2), 67–74. https://doi.org/10.1007/S00414-002-0352-5

Toscanini, U., Brisighelli, F., Moreno, F., Pantoja-Astudillo, J. A., Morales, E. A., Bustos, P., Pardo-Seco, J., & Salas, A. (2016). Analysis of Y-chromosome STRs in Chile confirms an extensive introgression of European male lineages in urban populations. *Forensic Science International. Genetics*, *21*, 76–80. https://doi.org/10.1016/J.FSIGEN.2015.12.005

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. *International Journal of Legal Medicine*, *107*(1), 13–20. https://doi.org/10.1007/BF01247268

Valdes, A. M., Slatkin, M., & Freimer, N. B. (1993). Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics*, *133*(3), 737–749. https://doi.org/10.1093/GENETICS/133.3.737

Vicard, P., & Dawid, A. P. (2004). A statistical treatment of biases affecting the estimation of mutation rates. *Mutation Research*, *547*(1–2), 19–33. https://doi.org/10.1016/J.MRFMMM.2003.11.005

Vicard, P., Dawid, A. P., Mortera, J., & Lauritzen, S. L. (2008). Estimating mutation rates from paternity casework. *Forensic Science International. Genetics*, *2*(1), 9–18. https://doi.org/10.1016/J.FSIGEN.2007.07.002

Vieira, T. C., Duarte Gigonzac, M. A., Goulart Rodovalho, R., Morais Cavalcanti, L., Bernardes Minasi, L., Melo Rodrigues, F., & da Cruz, A. D. (2017). Mutation rates in 21 autosomal short tandem repeat loci in a population from Goiás, Brazil. *Electrophoresis*, *38*(21), 2791–2794. https://doi.org/10.1002/ELPS.201700192

Weber, J. L., David, D., Heil, J., Fan, Y., Zhao, C., & Marth, G. (2002). Human diallelic insertion/deletion polymorphisms. *American Journal of Human Genetics*, *71*(4), 854–862. https://doi.org/10.1086/342727

Weber, J. L., & Wong, C. (1993). Mutation of human short tandem repeats. *Human Molecular Genetics*, *2*(8), 1123–1128. https://doi.org/10.1093/HMG/2.8.1123

Wierdl, M., Dominska, M., & Petes, T. D. (1997). Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics*, *146*(3), 769–779. https://doi.org/10.1093/GENETICS/146.3.769

Willems, T., Gymrek, M., Poznik, G. D., Tyler-Smith, C., & Erlich, Y. (2016). Population-Scale Sequencing Data Enable Precise Estimates of Y-STR Mutation Rates. *American Journal of Human Genetics*, *98*(5), 919–933. https://doi.org/10.1016/J.AJHG.2016.04.001

Wyman, A. R., & White, R. (1980). A highly polymorphic locus in human DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *77*(11), 6754–6758. https://doi.org/10.1073/PNAS.77.11.6754

Xavier, C., de la Puente, M., Sidstedt, M., Junker, K., Minawi, A., Unterländer, M., Chantrel, Y., Laurent, F. X., Delest, A., Hohoff, C., Bastisch, I., Hedman, J., van der Gaag, K. J., Sijen, T., & Parson, W. (2022). Evaluation of the VISAGE basic tool for appearance and ancestry inference using ForenSeq® chemistry on the MiSeq FGx® system. *Forensic Science International: Genetics*, *58*, 102675. https://doi.org/10.1016/J.FSIGEN.2022.102675

Xu, X., Peng, M., Fang, Z., & Xu, X. (2000). The direction of microsatellite mutations is dependent upon allele length. *Nature Genetics*, *24*(4), 396–399. https://doi.org/10.1038/74238

Yang, Y., Wang, W., Cheng, F., Chen, M., Chen, T., Zhao, J., Chen, C., Shi, Y., Li, C., Chen, C., Liu, Y., & Yan, J. (2018). Haplotypic polymorphisms and mutation rate estimates of 22 Y-chromosome STRs in the Northern Chinese Han father–son pairs. *Scientific Reports 2018 8:1*, *8*(1), 1–6. https://doi.org/10.1038/s41598-018-25362-3

2. Aims and Results

The accurate estimation of STRs mutation rates is of the utmost importance for the correct analysis, interpretation, and quantification of experimental data in a wide range of problems in anthropology, evolution, and population and forensic genetics. Since mutations are rare events, there is the need for a substantial amount of data to attain narrower confidence intervals and, thus, more accurate mutation rates and modelling. The accurate estimation of STRs mutation rates is, however, a complex endeavor as the most used methodology to obtain STRs mutation rates is the Mendelian incompatibilities approach (MIA) applied to data generated through genotyping platforms with automated length determination after capillary electrophoresis (CE). The methodology relies on the observation of Mendelian incompatibilities between parent(s) and child duos or trios. However, only in the case of simple structured STRs in the haploid system (MSY – male specific region of the Y chromosome) it is possible to infer undoubtedly which paternal allele originated which filial one. For the haplodiploid and diploid systems (X chromosome and autosomes, respectively) the estimation of mutation rates through the MIA implies biases. Several of these biases were already studied, with corrections proposed for autosomal STRs.

Hence, in this work, our aims were:

- To improve the estimation of STRs mutation rates through the revision of literature, and the production and collection of new data using exclusively one-generation familial pedigrees for both X- and Y-chromosome markers - Section 2.1. STR mutation rates estimation;

- To study, describe and measure the different biases introduced by the (pseudo) most parsimonious reasoning used in the interpretation of allele incompatibilities in familial genotypic configurations, in terms of both the number of mutational steps and the parental origin of the mutation, and to propose corrections to account for the occurrence of hidden mutations for both autosomal and X-chromosomal STRs - Section 2.2. STR mutation rates estimation biases;

- To improve mutation modeling through the analysis of the different factors influencing the mutation rate of STRs - Section 2.3. STR mutation rates modeling.

2.1. STR mutation rates estimation

Since mutations in STRs are rare events, large databases are needed to properly estimate mutation rates and accomplish adequate confidence intervals. The gathering of published, peer-reviewed, results and production of new biological data is, thus, crucial and should be encouraged. The works presented in this section were developed with the aim of expanding the knowledge in this subject.

- To obtain a thorough outlook of the state of the art of the use of X chromosome markers in forensic and population genetics, a comprehensive review of the literature was performed. The review focused on the developments achieved mainly regarding: i) the number and types of markers available, ii) the existing worldwide population data, iii) the statistical evaluation of the evidence in kinship testing, iv) the segregation and mutation studies available and v) the current weaknesses and prospects of their use. This work resulted in a publication on Frontiers in Genetics: Gomes, I., *et al.* (2020). Twenty years later: a comprehensive review of the X chromosome use in forensic genetics. *Frontiers in Genetics*. (https://doi.org/10.3389/fgene.2020.00926) and is presented in section 2.1.1. of this thesis. Please note that due to space constraints supplementary Table 2 is only accessible online at https://doi.org/10.3389/fgene.2020.00926.

- Assuming a stepwise mutation model, STRs' tendency to expand or contract, i.e., gain or lose repeats, respectively, was ascertained for 19 Y-STRs and 104,083 allele transfers from a range of 104 father-son duos for marker DYS549 to 9,108 for marker DYS391, which resulted in a publication on Forensic Science International: Genetics Supplement Series: Antão-Sousa, S., *et al.* (2019). Mutation in Y STRs: Repeat motif gains vs. losses. *Forensic Science International: Genetics Supplement Series*. (https://doi.org/10.1016/j.fsigss.2019.09.092) and was presented as a conference poster at the 28th Congress of the International Society for Forensic Genetics (ISFG). This work is presented in section 2.1.2. of this thesis.

- The Spanish and Portuguese speaking group of the International Society for Forensic Genetics (GHEP-ISFG) organized a collaborative study for its members

with the goal to gather Y-STRs mutation rates data from father-son duos. The study was proposed during the 2019 General Assembly and the participation open to all GHEP members. The participants of this GHEP-ISFG working group managed to collect data from 2,225 father-son duos, from 10 populations. The duos were analyzed with either PowerPlex® Y23 System, Promega (PPY23), or Yfiler™ Plus PCR Amplification kit, ThermoFisher Scientific (YFPlus). Specifically, we analyzed 203 duos from residents of North Portugal with the PPY23 kit. GHEP-ISFG new data were gathered to data collected from 44 published works, corresponding to 25,729 duos. This work is under final preparation and will be submitted as: Antão-Sousa, S. & Gusmão, L., *et al.* (2023). Microsatellites' mutation modeling through the analysis of the Y-chromosome transmission: results of a GHEP-ISFG Collaborative Study. It is presented in section 2.1.3. of this thesis

- Another GHEP-ISFG collaborative study was proposed to study X-STRs mutation rates and populational data from father-mother-daughter trios. The trios were analyzed with either the Investigator® Decaplex SE kit, Qiagen, or Investigator® Argus X-12 QS kit, Qiagen. Specifically, we analyzed 137 trios from residents of North Portugal with the Argus X-12 QS Kit. The mutation data produced in the scope of the collaborative study were then added to the data retrieved from 20 publications and the results will be submitted as: Gusmão, L. & Antão-Sousa, S., *et al.* (2023). X-Chromosomal STRs: metapopulations and mutation rates. It is presented in section 2.1.4. of this thesis.

2.1.1. Twenty years later: a comprehensive review of the X chromosome use in forensic genetics

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*, 926. doi: 10.3389/fgene.2020.00926

For supplementary material please check Appendix A.

frontiers in Genetics

REVIEW
published: 17 September 2020
doi: 10.3389/fgene.2020.00926

Twenty Years Later: A Comprehensive Review of the X Chromosome Use in Forensic Genetics

*Iva Gomes[1,2], Nádia Pinto[1,2,3], Sofia Antão-Sousa[1,2,4,5], Verónica Gomes[1,2], Leonor Gusmão[5] and António Amorim[1,2,4]**

[1] Institute for Research and Innovation in Health Sciences (i3S), University of Porto, Porto, Portugal, [2] Institute of Molecular Pathology and Immunology, University of Porto (IPATIMUP), Porto, Portugal, [3] Center of Mathematics, Faculty of Sciences, University of Porto, Porto, Portugal, [4] Department of Biology, Faculty of Sciences, University of Porto, Porto, Portugal, [5] DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Rio de Janeiro, Brazil

OPEN ACCESS

Edited by:
Kenneth K. Kidd,
Yale University, United States

Reviewed by:
Guanglin He,
Sichuan University, China
Carlo Robino,
University of Turin, Italy

***Correspondence:**
António Amorim
aamorim@ipatimup.pt

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *31 May 2020*
Accepted: *24 July 2020*
Published: *17 September 2020*

Citation:
Gomes I, Pinto N, Antão-Sousa S, Gomes V, Gusmão L and Amorim A (2020) Twenty Years Later: A Comprehensive Review of the X Chromosome Use in Forensic Genetics. Front. Genet. 11:926. doi: 10.3389/fgene.2020.00926

The unique structure of the X chromosome shaped by evolution has led to the present gender-specific genetic differences, which are not shared by its counterpart, the Y chromosome, and neither by the autosomes. In males, recombination between the X and Y chromosomes is limited to the pseudoautosomal regions, PAR1 and PAR2; therefore, in males, the X chromosome is (almost) entirely transmitted to female offspring. On the other hand, the X chromosome is present in females with two copies that recombine along the whole chromosome during female meiosis and that is transmitted to both female and male descendants. These transmission characteristics, besides the obvious clinical impact (sex chromosome aneuploidies are extremely frequent), make the X chromosome an irreplaceable genetic tool for population genetic-based studies as well as for kinship and forensic investigations. In the early 2000s, the number of publications using X-chromosomal polymorphisms in forensic and population genetic applications increased steadily. However, nearly 20 years later, we observe a conspicuous decrease in the rate of these publications. In light of this observation, the main aim of this article is to provide a comprehensive review of the advances and applications of X-chromosomal markers in population and forensic genetics over the last two decades. The foremost relevant topics are addressed as: (i) developments concerning the number and types of markers available, with special emphasis on short tandem repeat (STR) polymorphisms (STR nomenclatures and practical concerns); (ii) overview of worldwide population (frequency) data; (iii) the use of X-chromosomal markers in (complex) kinship testing and the forensic statistical evaluation of evidence; (iv) segregation and mutation studies; and (v) current weaknesses and future prospects.

Keywords: X chromosome short tandem repeats (X-STRs), X chromosome markers, forensic genetics, population genetics, kinship testing, X chromosome short tandem repeat (X-STR) mutation rates

INTRODUCTION

The X chromosome has many characteristics that are not shared by its counterpart, the Y chromosome, or by any of the autosomes of the mammalian genome. Its unique structural characteristics have been shaped by evolution, leading to the present known gender-specific genetic differences (Lahn and Page, 1999; Schaffner, 2004). In males, the single copy of the X chromosome does not allow for recombination to occur (except for the pseudoautosomal regions, PARs, which maintain homology by recombining during male meiosis). The non-recombining regions and the PAR 1 and PAR 2 regions of the X and Y chromosomes have taken different evolutionary paths becoming highly differentiated due to different functional roles, and consequently, only a few X-Y sequence similarities remain among them (Lahn and Page, 1999). Mutation events have gathered on the Y chromosome, and in addition to the lack of recombination, these events have contributed to the loss of most of the Y chromosome's sequence and genes emerging in a distinctive configuration of repeated sequences (Lahn and Page, 1999; Schaffner, 2004) becoming specialized in male sex determination. On the other hand, the X chromosome has preserved its autosomal character, becoming one of the most stable nuclear chromosomes, holding the largest and most conserved gene arrangement across eutherian ("placental") mammals (Lahn and Page, 1999; Kohn et al., 2004; Schaffner, 2004). It is the only chromosome to have one of its pair inactivated in one sex (females), and it is "corrupted" with repeat elements, making it especially tough to produce a detailed gene sequence (Gunter, 2005). In 2005, Ross et al. (2005) published the first draft that covered approximately 99.3% of the human X chromosome euchromatic sequence. The X chromosome holds a size length of approximately 155 million base pairs (Mb) (Ross et al., 2005), representing nearly 5% of the estimated human genome size (3,200 Mb) (Lander et al., 2001). Regarding some of the X chromosome's structural properties, it presents a low GC content (39%) when compared to 41% of the genome average (Ross et al., 2005). The low number of functional genes detected confers the chromosome one of the lowest gene densities among the chromosomes annotated to date (Ross et al., 2005). The X chromosome's sequence data revealed not only a low concentration of genes but also small gene length as only 1.7% of the chromosome sequence is represented by exons of the identified genes, responsible for transcribing 33% of the X chromosome (Ross et al., 2005). The particular genetic characteristics of the X chromosome, shaped by evolution, are responsible for the distinctive gender-specific features (**Figure 1**): in the male gender, the X chromosome is (almost entirely) transmitted to females as an unchanged block. While in females, the two copies recombine, like autosomes, reorganizing genetic variation in each generation, which contributes to the increase of haplotype diversity (Schaffner, 2004). The new reshuffled chromosome is then transmitted to female and male descendants (**Figure 1**).

These specific properties – two recombining copies in females and a single non-recombining copy in males (except for the PAR regions) creating haplotypes – provide X chromosome markers a particular place in forensics and in population genetics, as well as in other research areas such as human evolutionary studies and medical genetics (e.g., X-linked recessive disorders such as hemophilia or Duchenne muscular dystrophy) (Szibor, 2007). Regarding forensic and population genetic applications, the X chromosome's mode of inheritance places this chromosome among the autosomes and the uniparental-inherited genomes [mitochondrial DNA (mtDNA) and Y chromosome] providing desirable and exclusive features that are not provided by any other of the latter.

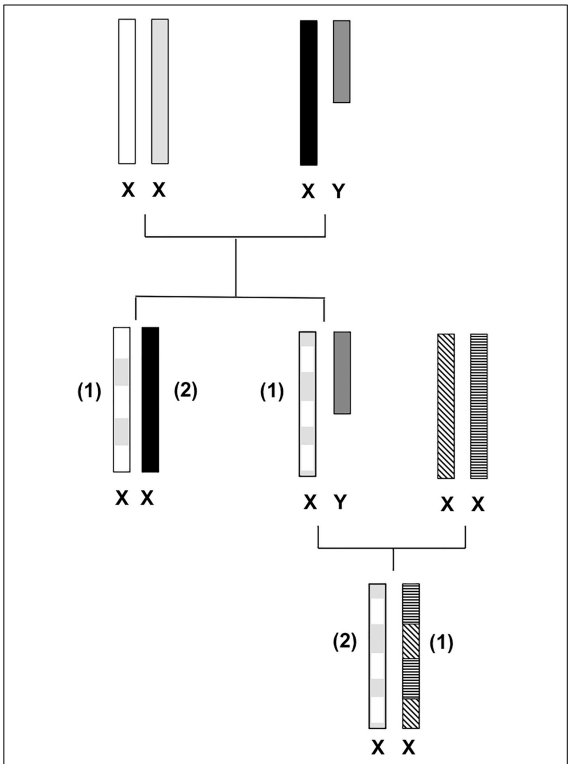


FIGURE 1 | X-chromosomal inheritance. Female and male descendants inherit a recombined maternal X chromosome (1) that resulted from female meiosis. Female offspring inherit one paternal unchanged X chromosome (2) due to lack of recombination [with exception for the pseudoautosomal region (PAR) regions].

In the early 2000s, the number of publications using X-chromosomal polymorphisms in these areas of research increased steadily. However, nearly 20 years later, a conspicuous decrease in the rate of these publications is observed. For example, X chromosome short tandem repeat (X-STR) forensic-based publications reached as many as 43 publications in a single year (2009), while in the past year of 2019, only 18 publications were found (complete results and detailed information on the criteria used for database search are presented and discussed under the section "Factors Underlying the Relative Stagnation

in X Chromosome Forensic Research"). In light of these observations, the main aim of the present work is to provide an up-to-date and objective review of the advances and applications of X-chromosomal markers in population and forensic genetics over the last two decades since the bloom observed in the early and mid-2000s.

CURRENT DEVELOPMENTS: NUMBERS AND TYPES OF X-CHROMOSOMAL MARKERS AVAILABLE (SHORT TANDEM REPEATS, SINGLE-NUCLEOTIDE POLYMORPHISMS, AND INSERTIONS/DELETIONS)

The use of X chromosome polymorphisms in human identification and in population genetics is mainly supported by the potential applications that outcome from its unique properties. Solely, or complementing the information provided by the autosomes or by markers located on the Y chromosome or mtDNA, X chromosome markers may provide essential information in many different lines of research. It must be highlighted that identity testing using X-STRs in particular contexts, namely in scenarios of (complex) kinship testing, may be the only tool to unravel certain cases. Examples of complex kinship testing scenarios where the prominent role of X chromosome polymorphisms is demonstrated are given in the section "The Use of X-Chromosomal Markers in (Complex) Kinship Testing."

In the present section, we will try to draw the state of the art of the genetic markers that have been described, to date, in the X chromosome-specific region, i.e., leaving out PARs and amelogenin. Special attention is given to X-STRs as a result of their favorite usage in forensic genetics due to high standardization and existence of commercial typing kits. Although some of the first publications reporting X-STRs appeared in the late 90s (Edwards et al., 1991, 1992; Hearne and Todd, 1991; Sleddens et al., 1992), the beginning of the century marked the increase of X-STR publications that focused on the development of new multiplexes on the genetic characterization of many different population groups (databasing) and on kinship and forensic investigations.

An extensive literature review was undertaken, with special focus on forensic-population genetic publications. Results are analyzed and tabulated separately for each type of marker, including relevant references. **Supplementary Table 1** lists 85 STR loci in which usage in forensic-population genetic context was reported. In agreement with the study of Szibor et al. (2005), HumARA marker was not considered for ethical reasons. Although the number of X chromosome markers has grown since the 2007 seminal review of Szibor (2007), this growth may be illusory, since many markers were used quite rarely, sometimes only once.

Although a considerable number of X-STRs are available in the literature, a better view of their real, current usage may be given by the analysis of the multiplexes, which have been described for their genotyping. **Table 1** shows the most used in-house and commercially developed X-STR multiplexes in which we update the revision of Diegoli (2015) and demonstrate clearly that the effective number of STRs routinely used is modest.

In any case, due to their high degree of discrimination, the number of standardized STRs is sufficient for most routine investigations, as will be discussed below in the section "The Use of X-Chromosomal Markers in (Complex) Kinship Testing." Novel interesting STRs for forensic applications continue being described (Nishi et al., 2020). Despite the wide set of available X-STR markers as well as many population-based studies (see section "Overview of Worldwide (Published) X Chromosome Short Tandem Repeat Population Data") that have emerged over these years, no effective X-STR database exists harboring this type of data. Some of the published population datasets are available in the FamlinkX web page[1] in a format that can be directly uploaded for kinship calculation using the software. Efforts were made by Szibor et al. (2006) to create an X-STR database[2] (ChrX-Str.org 2.0, 2020) that could anchor population data (namely, haplotype frequencies), calculation of forensically relevant parameters, information on markers such as multiplex kits, etc. Nevertheless, it seems that no updates have been made to this database, specifically in what regards population data submission, as only four populations are currently available (German, Ghanesen, Japanese, and Chinese Han) ("Haplotypes"; see text footnote 2). In addition, it is however noteworthy that no autosomal STR database such as NIST STRbase (National Institute of Standards, and Technology [Nist], 2020) or STRidER (2020)[3] contains information on X-STRs either. This approach could be considered: autosomal types of database could potentially serve as harbor for X-STR data undergoing the same quality control (QC) submission criteria. In fact, several forensic-focused journals such as the *Forensic Science International: Genetics* and the *International Journal of Legal Medicine* have published minimum requirements for publication of forensic population data from different genomic markers (e.g., autosomal, Y-chromosomal, mtDNA) (Parson and Roewer, 2010; Gusmão et al., 2017). Submission of such data to these journals requires preliminary QC assessment and inclusion in public online databases. These requirements could certainly be applied to X-chromosomal type of markers, ensuring the same quality type of data submitted. STRs are undoubtedly the preferential markers in human identification applications. Some of the main features that make STRs desirable markers are (i) highly polymorphic, i.e., high discriminating capacity between individuals; (ii) technical easiness due to rapid analysis with PCR-based technology and capillary electrophoresis automated fluorescent detection; and (iii) ability for generating STR multiplexes with small amplicon lengths for degraded DNA. The same cannot be said about insertions/deletions (INDELs), although these share some of the features of STRs (technical ease of analyses by PCR and ability for multiplexing), standardization is much less advanced perhaps due to the

[1] http://famlink.se/Databases/
[2] http://www.chrx-str.org/
[3] https://strider.online/

TABLE 1 | Most used multiplex PCR assays targeting X chromosome short tandem repeat (STR) markers. References do not necessarily refer to the original development papers.

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)

need of a much higher number of markers for a high degree of discrimination among individuals. Nevertheless, INDELs represent another potential tool for addressing human genetic identification issues. In **Table 2**, we list the X chromosome-specific INDEL polymorphisms genotyping systems described in forensic literature.

Unsurprisingly, not as many X chromosome INDEL marker systems have been described as compared to autosomal INDELs (e.g., Pereira et al., 2009; Freitas et al., 2010; Zaumsegel et al., 2013). In fact, no commercial kits being available, few systems have been subject to interlaboratorial comparisons, as in the case of autosomal INDELs, which stood international collaborative exercises (Pereira et al., 2018). One of the possible motifs for the lack of commercial kits is possibly due to the limited applications of X chromosome polymorphisms in forensic genetics when compared to autosomal markers. An interesting alternative typing approach, however, albeit of difficult analysis, is the one described in the studies of Fan et al. (2015, 2016) in which amplicons comprise various INDELs, i.e., biallelic loci that are tightly linked composing a new marker and that are amplified by a single pair of PCR primers.

TABLE 2 | X chromosome specific insertion/deletion (INDEL) polymorphisms genotyping systems. CE, capillary electrophoresis.

Number of loci	Genotyping system	References
32	Single multiplex (CE)	Pereira et al. (2012)
33	Single multiplex (CE)	Freitas et al. (2010)
16 (from a total of 45 mixed marker system)	Single multiplex (CE)	Tao et al. (2019)
17 (from a total of 60 mixed STR system)	Massive Parallel Sequencing	Zhang et al. (2017b)
21	Single multiplex (CE)	Edelmann et al. (2016)

With respect to X chromosome single nucleotide polymorphisms (X-SNPs), the analysis of the state of the art is even more complex due to the diversity of non-standardized genotyping systems and platforms, which have not been submitted to interlaboratorial comparisons. In **Table 3**, a summary of the actual forensic use of X chromosome-specific SNPs is shown. The number of table entries gives a false impression of abundance of X-SNPs; in fact, besides the

TABLE 3 | X chromosome-specific single-nucleotide polymorphism (SNP) genotyping systems.

Number of SNPs	Genotyping system	References
28 (from a total of 60 mixed marker systems)	Massive parallel sequencing	Zhang et al. (2017b)
39 (from a total of 273 mixed marker panels)	Massive parallel sequencing	Zhang et al. (2017a)
27 (from a total of 1,204 mixed marker panels)	Massive parallel sequencing	Hwa et al. (2018)
62	MALDI-TOF mass spectrometry	Stepanov et al. (2016)
17 (from a total of 220 mixed marker panels)	MALDI-TOF mass spectrometry	Hwa et al. (2019)
5 (from a total of 41 mixed marker panels)	MALDI-TOF mass spectrometry	Petkovski et al. (2005)
10	qPCR (TaqMan probes)	Zarrabeitia et al. (2007)
25	SNaPshot	Tomas et al. (2010)
16	SNaPshot	Oki et al. (2012)
14	qPCR (Taqman probes)	Li et al. (2010)

mentioned limitations, the number of SNPs overlap is very low. Although the binary nature of SNPs may favor degraded DNA as well as automation and high-throughput genotyping (e.g., in individual identification using complex kinship analyses in highly degraded scenarios such as natural or human-made disasters), the information content is considerably lower than for STR loci and consequently a larger number of SNPs are needed to match the discrimination power of the commonly used STRs (e.g., Chakraborty et al., 1999; Amorim and Pereira, 2005). Consequently, more loci mean more amplification products, which increases difficulty in data interpretation of DNA profile mixtures. In a multiple-donor sample interpretation, identification of each contributor may be very complex with biallelic systems. The limited number of alleles for each locus (normally two alleles) becomes hard to interpret because overlap will occur and multiple donors become hard to distinguish (Butler et al., 2007; Budowle and van Daal, 2008). Adding the mentioned data interpretation complexity in mixed profiles to the limited applications of X chromosome markers can potentially justify the lack of interest in X-SNPs observed.

OVERVIEW OF WORLDWIDE (PUBLISHED) X CHROMOSOME SHORT TANDEM REPEAT POPULATION DATA

For an overview of the worldwide population allele frequency datasets of X-STRs used in forensic genetics, we have consulted the articles available in PubMed database and in the congress proceedings of the International Society for Forensic Genetics[4] (The International Society for Forensic Genetics [ISFG], 2020). This search resulted in a total of 269 articles. The first genetic studies with focus on genotyping X-STRs for forensic application start emerging in the year 1999. Since then, and until 2008, a remarkable increase of population data publications was observed (**Figure 2A**). Nevertheless, reported information on human X-STRs in different worldwide populations has been stagnating in the last years.

Information concerning the populations, number of male and female samples, and X-STRs analyzed was compiled using 236 publications out of the 269 consulted (see **Supplementary Table 2**). The remaining were excluded for different reasons, which include articles that were not in English, with overlapping data (in this case, the most updated dataset was considered), and with unclear information concerning population, markers, or total samples analyzed. Therefore and although some of these studies contain relevant information on X-STR variation (e.g., the study by Edelmann et al., 2006, which has data for DXS9908 and DXS7127 markers), these were not included in **Supplementary Table 2**. Furthermore, the study by Phillips et al. (2018) reports data on seven X-STRs for a large sample of 944 individuals from the HGDP-CEPH human genome diversity panel. However, since this dataset comprises samples from 51 populations with relatively low sample sizes, the results were compiled for seven continentally defined population groups, namely, African (sub-Saharan), European, Middle East (including North Africans), Central-South Asian, East Asian, Oceanian, and Native American.

In **Figure 2B**, it is possible to observe that the number of X-STRs analyzed is highly variable among publications with some studies genotyping a high number of X-STRs (e.g., Liu et al., 2013; Fukuta et al., 2019) and others genotyping a reduced number of loci (e.g., Watanabe et al., 2000; Koyama et al., 2002; Carvalho and Pinheiro, 2011). The number of markers included in each study varied from 1 to 27. In 47% of the cases, this number was between 10 and 12 X-STRs, in 31%, it was below 10, and 22% of the datasets included more than 12 makers (**Figure 2B**). The number of markers available per dataset is somehow related to the use of commercial kits in 37.4% of the population studies (**Supplementary Table 2**). The first commercial kit that was optimized for forensic applications was the Argus X-UL from Biotype (Dresden, Germany), containing four X-STRs (DXS8378, DXS7132, HPRTB, and DXS7423) located in distant positions along the chromosome to avoid linkage. This kit was soon expanded (Argus X-8) with four additional X-STRs (DXS10135, DXS10074, and DXS10134), creating four pairs of linked X-STRs. The Argus X-12 (Qiagen, Hilden, Germany) is the most recent version of the Argus kit and is the most widely used (an optimized version is now available, the Argus X-12 QS, but that contains the same markers). It comprises 12 X-STRs organized in four linkage groups: LG1, DXS10148/DXS10135/DXS8378; LG2, DXS7132/DXS10079/DXS10074; LG3, DXS10103/HPRTB /DXS10101; and LG4, DXS10146/DXS10134/DXS7423. The Goldeneye DNA ID System 17X (Goldeneye Technology Co., Ltd., Beijing, China) and the AGCU X19 STR kit (Wuxi Sino-German Meilian Biotechnology Co., Jiangsu, China) were also developed for forensic applications, although available data are virtually restricted to Chinese populations. Among in-house

[4]www.isfg.org

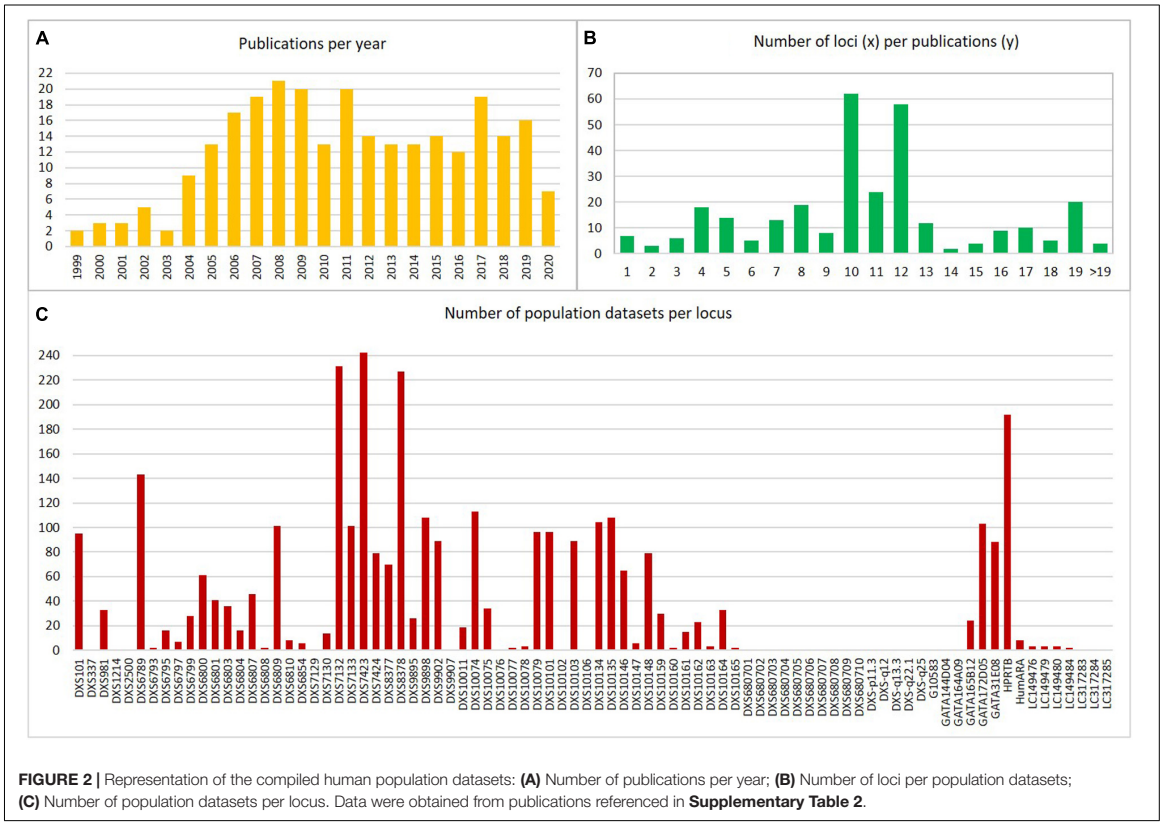


FIGURE 2 | Representation of the compiled human population datasets: **(A)** Number of publications per year; **(B)** Number of loci per population datasets; **(C)** Number of population datasets per locus. Data were obtained from publications referenced in **Supplementary Table 2**.

multiplexes, the Decaplex system developed by the GHEP-ISFG (Spanish and Portuguese Speaking working group of the ISFG) (Gusmão et al., 2008) has been the most widely used (14.6% of the population datasets were generated using this multiplex).

From the 84 markers that have been described as informative for forensic applications [including HumARA that is no longer used due to ethical issues (Szibor et al., 2005), as already mentioned], less than 50% were studied in more than 10 populations, and 29 were only reported in a single population (**Figure 2C**). The loci with more allele frequency data accumulated are those included in the commercial kits (namely, Investigator Argus X-12 kit, Qiagen) or in the in-house-developed Decaplex-GHEP-ISFG (Gusmão et al., 2008).

In **Supplementary Table 2**, the geographical distribution of the published human population data for X-STRs since 1999 is described. Notwithstanding the exhaustive nature of this review, it is possible that some studies are missing from this table. However, we believe that most forensic population studies on X-STRs have been identified, allowing a realistic picture of the state of the art. For a broader overview of the populations sampled, we have represented the number of datasets that have been published until now by country (**Figure 3**). The datasets were counted considering the number of subpopulations or ethnic groups in each publication. Populations defined at continental level (namely, the HGDP-CEPH and Africa datasets) or belonging to ethnic affiliated populations from different countries (namely, the Jews) have been excluded. In **Figure 3**, it is possible to observe that apart from a lack of X-STR data information for many countries, there is high heterogeneity among and inside continents. Data are scarcer in some geographical areas, namely, for sub-Saharan African and American populations (except for Argentina, Brazil, and United States). On the other hand, a large quantity of X-STR data was obtained for other populations, such as the ones from China. China is by far the best represented country not only because of the higher number of publications but also due to the inclusion of various ethnic groups in a single study. Although for some countries a large number of datasets are available for the same X-STR loci, many of those studies characterize different regions or subpopulations, which is relevant to investigate population stratification inside the country, especially when a high diversity of ethnicities coexists.

Overall, the compiled information clearly shows an imbalance between the total number of publications and the asymmetric representation of the worldwide populations. In fact, for several populations from different geographic regions, data on X-STR remain largely scarce, being the available information representative of only a small fraction of the worldwide human populations. Moreover, apart from a large variation concerning the X-STRs included in each study, many only comprise a small

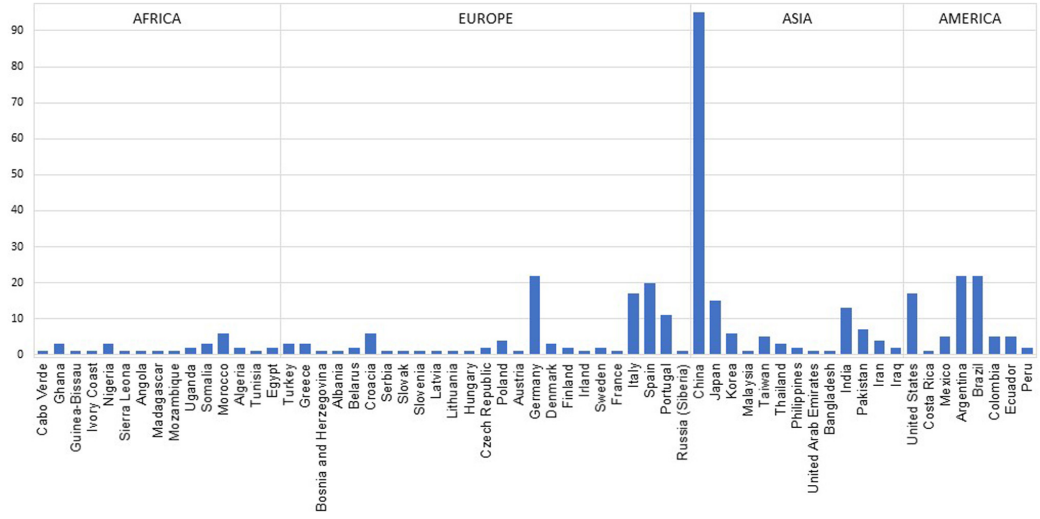


FIGURE 3 | Compilation of number of datasets per country. Data were obtained from publications referenced in **Supplementary Table 2**.

number of loci. Due to proximity on the chromosome, it is expected that some of the studied markers will be in linkage disequilibrium (LD) in many populations. However, data on haplotype frequencies are almost restricted to recent papers and not available for most publications consulted, invalidating the use of some of the available data in forensic applications.

Therefore, further studies on haplotype frequency distributions, as well as on mutation rates and LD, are mandatory to attain the final goal of establishing highly comprehensive and representative human reference X-STR databases.

SHORT TANDEM REPEAT NOMENCLATURES AND PRACTICAL CONCERNS

Accuracy and common nomenclature are of fundamental importance to secure error-free communication, data exchange, and data comparison among laboratories. STR nomenclature, independently of marker genome location, has been long addressed by several studies (e.g., Lazaruk et al., 2001; Gusmão et al., 2002; Gomes et al., 2008, 2009, 2016; Gettings et al., 2015) as well as by the ISFG and other DNA groups (e.g., Bär et al., 1997; Olaisen et al., 1998; Gill et al., 2001; Gusmão et al., 2006).

The observed increase of X-STR studies over the years justifies the need to evaluate X-STR nomenclature being used at least for the most common polymorphisms. Several studies have gathered considerable sequencing data for some of the commonly used X-STRs (Gomes et al., 2008, 2009, 2016, 2017; Szibor et al., 2009). In these latter studies, relevant findings were reported for several markers, which demonstrate that accurate allele nomenclature designation taking into consideration the ISFG recommendations (Gusmão et al., 2006) would have had a major impact on allele assignment. One of the major gaps seen in several studies is the lack of sequencing data for, at least, the three major population groups (Asian, African, and Caucasian) when new markers are proposed as usually only one group is analyzed. This approach reduces possible interpopulational variation and avoids genotyping problems when different groups are genotyped. This was the case for the first version of the most used X-STR commercial kit, the Investigator Argus X-12 (Qiagen). The markers in this kit were characterized mostly in individuals of European ancestry and therefore some of the genetic variations detected in other population groups were missed out (Tillmar et al., 2017). Once other population groups of other ancestries were studied, several markers presented high frequencies of silent alleles that had gone previously undetected (e.g., Tomas et al., 2012; Gomes et al., 2016, 2017; Tillmar et al., 2017). For example, the silent alleles for some of the loci were mostly caused by a mismatch at one of the primer binding sites (Gomes et al., 2016, 2017). After several reports on this matter, a new version was developed, the Investigator Argus X-12 QS (Qiagen), containing the same markers but with new primer designs for some of the X-STRs to resolve the high frequency of allele dropouts. Another example of inaccurate nomenclature assignment was the case of HPRTB. In the study of Pereira et al. (2007), peculiar results during population comparison analyses of a Northern Portuguese population sample with other European groups were found. These findings led to a deeper investigation, leading to the discovery of issues behind the HPRTB nomenclature (Szibor et al., 2009). In this latter report, authors described that two different nomenclatures were being used among the forensic genetic community, leading to a shift in allele frequencies and consequently errors in data resulting from population comparisons-based analyses (e.g., Pereira et al., 2007).

Finally, as proposed by the ISFG recommendations on the use of X-chromosome markers (Tillmar et al., 2017), the previous recommendations on allele nomenclature already recognized for autosomal and Y-chromosomal-specific markers (Bär et al., 1997; Gill et al., 2001; Gusmão et al., 2006) can also be applied to X-STRs without the need for particular changes. It seems that very few studies take these recommendations into thoughtful consideration and no real significant advances have been made in this field. Accuracy in sequence variation and repeat structure and nomenclature of X-STRs are empirical and pending issues in forensic and population genetics research that are still often neglected.

THE USE OF X-CHROMOSOMAL MARKERS IN (COMPLEX) KINSHIP TESTING

The standard procedure to quantify the genetic evidence in kinship analyses relies upon independent autosomal markers and is grounded in Bayes' theorem. Typically, equal priors are considered, and a likelihood ratio (LR) comparing the probability of the observations assuming a pair of alternative, mutually exclusive, kinship hypotheses is computed (Gjertson et al., 2007). Indeed, autosomal information is the one generally considered, despite currently available X-chromosomal markers being able to provide great statistical power in some cases (Szibor et al., 2003; Krawczak, 2007; Szibor, 2007; Pinto et al., 2011a, 2013a; Gomes et al., 2012). From the set of the latter cases obviously excluded are those where there is a link "father–son" in both main and alternative hypotheses, as, for instance, in a "paternal grandfather–granddaughter" vs "unrelated" case analyzing a pair of individuals, as the first, when considering X-chromosomal transmission, equated to the second (Pinto et al., 2011a, 2012). In any case, the preference given to autosomal markers is easily justified and understood not only for allowing the same approach for each kinship problem, regardless of the sex of the involved individuals, but also because of independent transmission of the markers and, at least in most of the populations, absence of LD. Conversely, the analysis of X chromosome markers offers little room to consider only independently transmitted loci, and thus recombination rates and haplotype frequencies are in general required for statistical evaluation of the evidence.

Non-random association of alleles of different loci at a population-level LD (also known as gametic association) can result from population events like drift, selection, non-random mating, or admixture (Hedrick, 1987; Medina-Acosta, 2011). A close physical location of the markers, as well as population stratification, will influence the re-establishing of equilibrium. Consequently, LD results neither can be extrapolated from one population to another, nor are stable, even in a closed population, as recombination progressively breaks it. Moreover, haplotype frequencies cannot be inferred from allelic ones, and direct counting needs to be carried out.

Closely located markers are said to be in linkage if they are more prone to be inherited together, as a unit, than independently. Linkage between markers depends on chromosomal recombination rate (or frequency). Two markers are unlinked if recombination between them is expected to occur in each meiosis so that half of the gametic products would be recombinant and thus recombination fraction takes the value of 0.50. Obviously, linked markers are more prone to be in LD. Segregation analyses in one or multi generation family studies were performed, aiming to estimate recombination rates between X-STRs of interest through proper bioinformatic pipelines that take into account the possibility of mutation (Nothnagel et al., 2012; Diegoli et al., 2016; Bini et al., 2019), but population-based studies, as HapMap project (The International HapMap Consortium, 2007), can also be considered (Phillips et al., 2012). Mapping functions as Haldane's (Haldane, 1919) or Kosambi's (Kosambi, 1944) are used to convert genetic distances between markers in recombination rates. It is however noteworthy that in some kinship problems, as the one involving a pair of females and the hypotheses maternity and unrelated, the linkage is not needed to be taken into account as it cancels in the LR numerator and denominator (Tillmar et al., 2017). A general framework to understand in which case linkage has to be considered is still lacking, despite being known that disregarding it may lead to a significant over- or under-quantification of the genetic evidence (Tillmar et al., 2011; Kling et al., 2015b).

Contrarily to what occurs for autosomes, where a plethora of markers from 22 chromosomes can be chosen, linkage and LD are unavoidable issues in the case of X-chromosomal analysis. Due to the length of the X chromosome, a maximum of four unlinked X-STRs are estimated to be liable of being simultaneously analyzed. On the other hand, higher LD values are expected for X-chromosomal markers than for autosomes since recombination only occurs in female meioses, which have also smaller mutation rates than males (Shimmin et al., 1993; Schaffner, 2004). Finally, it should be noted that estimates of haplotype frequencies are not as accurate as the allelic ones since much larger databases are required: just considering a simple illustrative example, a set of three loci with 10 alleles each can potentially entail the estimation of 1,000 haplotype frequencies.

Few software packages are available for kinship evaluations considering X-chromosomal transmission, FamLinkX being the most relevant, taking into account the possibility of mutation, linkage, and LD (Tillmar et al., 2011; Kling et al., 2015a). Also, software to weigh the *a priori* power of a marker to exclude a claimed relationship was already developed (Egeland et al., 2014), and the ISFG recently provided general guidelines for using X-chromosomal markers in kinship testing (Tillmar et al., 2017).

Kinship Testing and the Identity-by-Descent Framework

Considering a number of generations beyond which individuals are assumed to be unrelated, kinship measurements are

based on the concept of identity-by-descent. Two alleles are called identical-by-descent (IBD) if they are copies of a given ancestral allele. Barring mutation, two alleles which are identical by descent must be therefore identical-by-state (IBS). For autosomal transmission, nine IBD partitions can be established considering the four alleles of a pair of individuals and their relationship (Jacquard, 1974; Weir et al., 2006; Pinto et al., 2010). This number reduces to three if non-inbred individuals are considered, likewise occurring for X-chromosomal transmission between a pair of females (Pinto et al., 2011a, 2012). Regarding X-chromosomal transmission, there are four IBD partitions involving a female–male pair (two if assuming a non-inbred female) and two for a pair of males (Pinto et al., 2011a). Independently of the mode of genetic transmission considered, the probabilities of the genotypic observations, assuming a specific hypothesis of kinship, depend on the IBD probabilities of the pedigree and on the frequency of the alleles (Weir et al., 2006; Pinto et al., 2011a). Pedigrees with the same IBD coefficients are said to belong to the same kinship class, as they are, theoretically, undistinguishable through the use of unlinked markers (Pinto et al., 2010, 2012). In **Table 4**, IBD probabilities are presented for a pair of non-inbred individuals considering autosomal and X-chromosomal modes of genetic transmission and a set of commonly analyzed relationships. Algebraic formulae for the probabilities of the observations, given the identity by descent partitions, can be found in Weir et al. (2006) and Pinto et al. (2010, 2011a, 2012), respectively, for autosomes and X-chromosomal markers. Finally, it should be noted that, assuming X-chromosomal mode of transmission, relationships are not symmetrical as probabilities of IBD sharing may differ. For example, while a pair of paternal aunt–nephew does not share X-IBD alleles (being thus equated to unrelated from the X-chromosomal point of view), a pair of paternal uncle–niece shares one pair of IBD alleles with 50% of chance.

Regardless of the mode of genetic transmission considered, striking statistical results could be obtained when the sharing of IBD alleles is mandatory, unless mutation occurs, for one of the two kinship hypotheses considered. For example, in a standard paternity problem ("unrelated" as alternative hypothesis), the probability of sharing a pair of IBD autosomal alleles (and thus IBS, barring mutation) is one, under the main hypothesis, and null under the alternative. In cases with daughters, this is also true for X-chromosomal markers, providing a higher *a priori* paternity exclusion power than autosomal ones (Krawczak, 2007; Pinto et al., 2013a).

In some cases, as in disaster victim identification problems, specific kinship hypotheses cannot be established, and a broader measure of kinship can be established to weigh the degree of relatedness before specifying more detailed hypotheses. In these cases, the coancestry coefficient, i.e., the probability of selecting two IBD alleles when each one is randomly chosen from each individual, can be computed. In this case, the analysis of the X chromosome can be of major importance as, in all the cases where transmission is not interrupted by a "father–son" link, the expected IBD sharing is at least the same as for autosomes – see **Table 5**, since no randomness is possible in the X-allele of a male. Coancestry coefficients can be estimated through the genotypes of the individuals (Pinto et al., 2011b, 2013b) and the combination of both types of genetic information can provide valuable insights on the genetic kinship linking the individuals.

Parenthood Testing

The X-chromosomal markers can be used to complement autosomal information when inconclusive or weak statistical results are achieved in standard parenthood testing where the alternative hypothesis is the individuals being unrelated. This can be due to the poor quality or low quantity of DNA in degraded samples, resulting in few analyzed markers or to other, more complex, situations where few Mendelian incompatibilities are found.

Compared with autosomes, X-chromosomal markers provide greater statistical power in trios, in paternity duos with daughters, and in maternity duos with sons. The X-chromosomal markers are not informative in paternity cases with sons, and for mother/daughter duos, the same statistical power is obtained for autosomal and X-chromosomal transmission.

When few Mendelian incompatibilities are found, this can be due to the alleged parent of the child being related to the true parent. A relatively common situation is the alleged father being either a full brother or the father of the true father of the child, in which case the probability of the alleged father and child sharing a pair of IBD alleles is 50%. In a paternity testing with a daughter, if the alleged father is a brother of the real one, the probability of uncle–niece sharing a pair of IBD X-alleles is also 50%. In all the other cases, this probability is null. Indeed, the analysis of X-chromosomal markers can be an efficient approach for excluding close relatives of the real father, unknowingly presented in a standard paternity case (Gomes et al., 2012).

Beyond Parenthood

In some cases, the alleged parent is not available for analysis, and sibship, or grandparenthood problems may emerge. In some of these cases, X-chromosomal markers can provide invaluable information, stronger than the one provided by autosomes. The most striking examples are those where the sharing of a pair of IBD X-alleles is mandatory. This occurs when the paternity of a daughter is questioned, being the alleged father unavailable for analysis, contrarily to his (unquestioned) mother or daughter. In both cases, the sharing of IBS alleles between analyzed females is mandatory for all the markers, unless mutation occurs. In these cases, the reached statistical power is the same for a paternity testing with autosomes when the alleged father is directly analyzed whether the mother of the child is available for analysis or not.

Another illustrating example is the kinship problem where the hypotheses are "full sisters" versus "unrelated."

TABLE 4 | Probability of two individuals sharing two, one, or no pairs of identical-by-descent (IBD) alleles, assuming a specific kinship for both autosomal (Aut) and X-chromosomal (X chr) modes of genetic transmission.

Pair of shared IBD alleles		**Two**		**One**		**None**	
Mode of transmission		**Aut**	**X chr**	**Aut**	**X chr**	**Aut**	**X chr**
Relationship				**Female–Female**			
Identical twins/Identity		1		0		0	
Mother–Daughter		0		1		0	
Full-Sisters		1/4	1/2	1/2	1/2	1/4	0
Grandmother–Granddaughter	Maternal	0	0	1/2	1/2	1/2	1/2
	Paternal				1		0
Aunt–Niece	Maternal				3/4		1/4
	Paternal				1/2		1/2
Half-sisters	Maternal				1/2		1/2
	Paternal				1		0
Unrelated			0	0		1	
				Female–Male			
Father–Daughter/Mother–Son		0	–	1		0	
Full brother–sister		1/4		1/2	1/2	1/4	1/2
Grandfather–Granddaughter/Grandmother–Grandson	Maternal	0		1/2	1/2	1/2	1/2
	Paternal				0		1
Uncle–Niece	Maternal				1/4		3/4
	Paternal				1/2		1/2
Aunt–Nephew	Maternal				3/4		1/4
	Paternal				0		1
Half-brother–sister	Maternal				1/2		1/2
	Paternal				0		1
Unrelated				0		1	
				Male–Male			
Identical twins/Identity		1	–	0	1	0	
Father-Son		0		1	0	0	1
Full-brothers		1/4		1/2	1/2	1/4	1/2
Grandfather–Grandson	Maternal	0		1/2	1/2	1/2	1/2
	Paternal				0		1
Uncle–Nephew	Maternal				1/4		3/4
	Paternal				0		1
Half-brothers	Maternal				1/2		1/2
	Paternal				0		1
Unrelated				0		1	

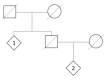
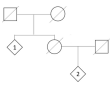
Considering X-chromosomal transmission and the main hypothesis, females share either two or one pair of IBD X-alleles with the same probability: 50%. Assuming autosomal transmission, they may not share IBD alleles (with 25% of chance), such as occurs assuming they are unrelated (with 100% of chance). It is then expected that X-chromosomal markers provide stronger results than autosomes. This occurs in all the kinships where the transmission of the X chromosome is not interrupted due to its obligatory transmission between father and daughter, which allows the skipping of one meiosis.

Incest Cases

In some cases, the high number of homozygosities shown by a child (e.g., in a paternity testing with alleged father excluded) may raise the suspicion of an incestuous situation. This may, under some circumstances, configure a crime (mother under age or with intellectual disability, for example). In the case of a daughter, X-chromosomal analyses may provide important insights even without analyzing the alleged father. If the father of the daughter is also the father of the mother and, in the absence of mutation, either the child is homozygous (for one allele present in the mother) or is heterozygous for the same alleles of the mother. In the case of autosomal transmission, three alleles can be seen in mother/daughter pair, as for the case of the parents being unrelated.

The hypotheses of the father of the child being either the father or the full brother of the mother are theoretically indistinguishable when considering unlinked autosomal markers. Contrastingly, in the case of daughters, X-chromosomal markers can provide insights allowing the different weighing of the two hypotheses (Pinto et al., 2011a).

TABLE 5 | Probability of choosing a pair of identical-by-descent (IBD) alleles when one allele is randomly chosen from each individual. Numbers in superscript in header refer to the sex of the individuals represented in genealogies.

Kinship		Coancestry	Female^1–Female^2	Male^1–Male^2	Female^1–Male^2	Male^1–Female^2
General		Aut-chr	$1/2k_2 + 1/4k_1$			
		X-chr	$1/2x_2 + 1/4x_1$	x_1	$1/2x_1$	$1/2x_1$
Parenthood		Aut-chr	1/4			
		X-chr	1/4	0	1/2	1/2
Full-sibship		Aut-chr	1/4			
		X-chr	3/8	1/2	1/4	1/4
Paternal half-sibship		Aut-chr	1/8			
		X-chr	1/4	0	0	0
Maternal half-sibship		Aut-chr	1/8			
		X-chr	1/8	1/2	1/4	1/4
Paternal grandparenthood		Aut-chr	1/8			
		X-chr	1/4	0	0	0
Maternal grandparenthood		Aut-chr	1/8			
		X-chr	1/8	1/2	1/4	1/4
Paternal avuncular		Aut-chr	1/8			
		X-chr	1/8	0	0	1/4
Maternal avuncular		Aut-chr	1/8			
		X-chr	3/16	1/4	3/8	1/8

k_i, probability of sharing i pairs of IBD autosomal alleles; x_i, probability of sharing i pairs of IBD X-chromosomal alleles.

Distinguishing Pedigrees Belonging to the Same Autosomal Kinship Class

Pedigrees are theoretically indistinguishable, considering unlinked markers, whenever they have the same IBD partitions (Pinto et al., 2010). This is the case of the second-degree relatives: avuncular, half-siblings and grandparent–grandchild, as the probability of individuals sharing two pairs of IBD alleles is null, while the probability of sharing one pair of IBD autosomal alleles is equal to the probability of sharing none (50%) – see **Table 4**. Nevertheless, the analysis of X-chromosomal markers can provide differential weighing favoring one of the alternative hypotheses (Pinto et al., 2011a). For example, when a pair of females is analyzed, maternal and paternal aunt/niece can be distinguished from, respectively, maternal and paternal half-sisters and grandmother–granddaughter, which are not distinguishable among them even when considering X-chromosomal markers. In all the cases, females cannot share two pairs of IBD alleles, but a pair of maternal aunt/niece shares one pair of IBD alleles with a probability equal to 75%, while for both maternal half-sisters and grandmother–granddaughter pairs, this probability reduces to 50%. On the other, if both pairs of paternal half-sisters and grandmother–granddaughter have to share one pair of IBD alleles, this probability drops from 100 to 50% in the case of paternal aunt/niece. Different IBD probabilities will result in different weighing of the evidence, depending on the genotypic observations.

SEGREGATION STUDIES: CURRENT DATA AND MISSING DATA

The high power of discrimination that characterizes STRs and makes them desirable genetic markers compared to SNPs or INDELs, particularly in human identification analysis (such as kinship testing), is due to their higher mutation rate. An STR is, by definition, a tandemly arrayed repetition of a DNA fragment of one to six base pairs. There is general consensus that these are created by random mutations (Levinson and Gutman, 1987; Schlötterer, 2000). Generally, STRs with four base pairs motifs are plentiful and more stable than two or three nucleotide repeats; hence, they have been favored when designing the commercially available forensic kits (Pereira and Gusmão, 2016). Motifs with two or three base pairs are less stable and have a higher propensity for stutter during PCR, and STRs with more base pairs are less frequent. When a somatic mutation occurs, it affects only cell lines of the individual where it occurred. However, when a mutation occurs in the germ line, it has the potential of being passed on to the offspring and resulting in different parental and filial alleles. Mutation rates vary between types of polymorphisms and also on inherent individual characteristics such as sex and age (Brinkmann et al., 1998; Nachman and Crowell, 2000).

Polymerase template slippage is thought to be the primary mutational mechanism leading to changes in STR length (Schlötterer and Tautz, 1992; Strand et al., 1993), and mutations involving the loss or gain of one repeat are assumed to be preponderant over mutations involving the loss or gain of multiple repeats. Slippage occurs during DNA replication when the two DNA strands come apart. When misalignment occurs out of register the repeat number of the STR product will be different. The currently accepted mutational model, also known as the stepwise mutation model (SMM) (Ohta and Kimura, 1973) occurring as a result of DNA replication slippage, includes mutational forces working in opposite directions: polymerase template slippage and point mutations; the latter reduce the length of STRs due to the breakage of the original segment creating two new shorter segments. Studies have shown that the longer the allele length, the higher is the frequency of these events. It has also been reported that longer alleles tend to mutate to shorter alleles and vice versa, while intermediate-sized alleles have approximately the same tendency to shorten or lengthen (Primmer et al., 1996; Brinkmann et al., 1998; Xu et al., 2000; Antão-Sousa et al., 2019).

In forensic casework context, the estimation of mutation rates is crucial for the analysis, interpretation, and quantification of experimental data and for the proper quantification of LRs. In such scenarios, the detection of mutation(s) has practical consequences in the interpretation of the genetic profiles. Some studies have addressed this by analyzing different familial configurations, familial duos, mother–son, mother–daughter, and father–daughter, and familial trios, father–mother–daughter (e.g., Jin et al., 2016; Burgos et al., 2019; García et al., 2019). **Supplementary Table 3** presents the most updated information on mutation rates per marker and per familial configuration for the most commonly used X-STRs. To date, not much research on the mutation rates of the most commonly used X-STRs has been given, and therefore, data collection and analyses are still lacking. Perhaps one of the limitations in the estimation of mutation rates of STRs, in general, is the use of the (most frequently used) method for mutation estimates based on direct pedigree analysis. This means that mutated alleles are identified straightforward by the observation of allele transmissions in parent–child requiring a large amount of data to reliably estimate allele mutation rates. Having access to a high number of specific constellations of families may be a drawback to the (accurate) estimation of mutation rates of X-STRs.

DISCUSSION

Factors Underlying the Relative Stagnation in X Chromosome Forensic Research

After an initial boom, forensic research interest on X chromosome markers has witnessed a decline as judged by the number of relevant publications: 2000 (6), 2001 (7), 2002 (11), 2003 (18), 2004 (27), 2005 (25), 2006 (40), 2007 (35), 2008 (42), 2009 (43), 2010 (19), 2011 (41), 2012 (26), 2013 (22), 2014 (18), 2015 (15), 2016 (22), 2017 (31); 2018 (16), and 2019 (18) [search results obtained using Scopus database[5] and the following criteria [ALL (dxs*) AND ALL (forensic)] AND PUBYEAR > 1999 AND PUBYEAR < 2020 on 30/04/2020]. In the beginning of the early 2000s, only a scarce number of X chromosome STRs and a very limited number of human population groups were characterized for forensic genetic applications. Data focusing on the assessment of X-linked polymorphisms for forensic and kinship genetic studies were an impending demand which created a gap in these fields producing sufficient ground for the interest in X chromosome markers and, in particular, X-STRs. Consequently, an increase of studies in 2003 until 2011 (with exception of the year 2010) can be noted. After this year, fluctuations are mostly toward a reduction of X-STR studies (except for 2017).

This implies that the practical forensic use of X chromosome is well below its potential and – what is most concerning – is that its use may be unsupported by research data and based on inadequately validated technical means and theoretically reduced or even incorrect analytical approaches. Enabling corrective actions demands therefore the identification of the causes of this slowing down of the forensically inclined research on X genetic markers. This fact has no parallel on the other sexual chromosome counterpart, the Y, to which a lot of attention is devoted, for example, by the STRbase (National Institute of Standards, and Technology [Nist], 2020) and has as well a very active dedicated site[6], YHRD (2020), in contrast to the ChrX-STR.org 2, as mentioned previously (see text footnote 2).

In this section, we will analyze the putative change counteracting the loss of interest and analyzing the presumed reasons or factors justifying this situation,

[5]http://www.scopus.com
[6]https://yhrd.org/

which, from our point of view, can be classified into four broad categories: (a) theoretical and/or analytical, (b) technical, (c) statistical, and (d) medical/ethical, to be detailed below.

Theoretical and Analytical Difficulties

The main obstacle to the correct use of X chromosome in forensics lies in the hybrid nature of its formal genetic model of inheritance, common to most mammals, with very few exceptions (Cortez et al., 2014; Matveevsky et al., 2017). Indeed, as presented in the section "Introduction," this chromosome harbors two distinct modes of transmission: the diploid, autosomal style (corresponding to the so-called pseudoautosomal regions), two in humans, PAR 1 and PAR 2 (Flaquer et al., 2009) and the sex-linked haplodiploid (for the rest of the chromosome, known as X-specific), which, due to the single copy in males, does not recombine.

When addressing X-chromosome markers, we are referring to the X-specific located ones. Therefore, only these will be analyzed (although some confusions do sometimes arise and quite often the status of X specificity may be doubtful – see below the technical section).

Even so, the formal genetic model of transmission and the consequences at the level of population genetics seem to be poorly understood by the forensic community, as judged by a recent analysis of the literature (Ferragut et al., 2019). It was shown that in 60% of 52 analyzed publications, forensic parameters were computed as for autosomal markers, and the analysis of associations between alleles from distinct loci (LD) was generally deficient or erroneous. In fact, linkage and LD concepts, particularly important for the X chromosome since all markers are located on the same chromosome, are often a source of confusion and generally lead to misinterpretation or even non-consideration of LD results in many genetic studies. Most studies using X-STRs correctly test for the presence of significant association among pairs of loci (LD) but fail to estimate haplotype frequencies and probability calculations, accordingly, when significant association is found among markers as loci must be analyzed together and not as individual markers in such cases. In 2017, recommendations were provided by the DNA commission of the ISFG addressing exactly the issues behind the concepts of linkage and LD in cases of kinship testing using X-STRs and emphasizing that "Haplotype frequencies should be used for likelihood calculations when LD exists" (Tillmar et al., 2017).

Similar issues have also arisen with the assessment of conformity with Hardy–Weinberg equilibrium expectations. Quite symptomatically, the ChrX-STR.org 2 website (see text footnote 2, accessed on 02/05/2020) has posted: "Based on the review of December 2018, it has been decided in cooperation with the X working group to remove the PI calculation from this website."

From an applicable point of view, one can add that one of the additional problems to justify the decrease of interest in X chromosome markers could be due to the low number of identification cases that request X-STR markers. Perhaps the troubles behind the implementation of a new system (financial cost and human resource training) which has a much more complex type of analysis when compared to the Y chromosome, for example, may not justify the need for the use of this system.

Technical Problems

Besides the genotyping problems, which may be transversal to all markers, irrespectively of the mode of transmission, sex chromosomes pose special difficulties due to their complex evolutionary history. In fact, apart from the PAR regions, X and Y chromosomes still keep substantial extensions of homologous regions, which obstruct the safe establishment of specificity for a marker, as well as its primers in case of PCR-based techniques. Particularly for recently X/Y transposed regions, this may constitute an (nearly) insurmountable obstacle (Lopes et al., 2004) as well as the dynamic state of the pseudoautosomal moving boundaries (Otto et al., 2011).

Statistical Issues

Most of the statistical problems (both at the descriptive level – parameter estimation level or hypothesis testing design or evidence quantitative evaluation) stem out of the theoretical flaws discussed above. Nonetheless, some are specifically empirical and are related to the haplodiploid specificity of the X chromosome: different sampling and estimation methods are required for each sex. Indeed, while haplotype frequencies can be estimated by simple counting in males, in females, they have to be inferred. Needless to say, simple haplotype frequency estimation requires prohibitively large sample sizes, growing exponentially with the number of loci involved (Amorim and Pinto, 2018).

Medical/Ethical Questions

To begin with, it must be highlighted that the very genotyping of sex chromosome markers for forensic purposes may represent a violation of some of the established recommendations and rules on the exclusion of any markers that can reveal physical traits [e.g., European Council Resolution of 25 June 2001 on the exchange of DNA analysis results (2001/C 187/01)]. Furthermore, gender and sex are always sensitive, and sometimes conflicting, categories in or for some individuals.

The evolutionary dynamics of sex chromosomes introduces also undesirable clinical and ethical problems. In fact, sex chromosomes are the Achilles' heel of male meiosis (Kauppi et al., 2012). A non-negligible proportion (1/448 live births) of the human population carries some sort of chromosomal aberration and, for example, one of the aneuploidies, Klinefelter syndrome, has an incidence of ∼1/500 male live births) (Nielsen and Wohlert, 1990). The consequences for forensic practice are ethically troublesome: discordance of external sex from X chromosome typing and unwilling disclosure of a clinical condition. In addition to X-chromosomal changes, several X-STR markers, that were or are still in use, have been linked to medical conditions. The HumARA is linked to spinal and bulbar muscular dystrophy (SBMA) as well as to other health risks

(Szibor et al., 2005). Another example is the possible LD between the STR alleles at HPRTB locus to the X-linked recessive disorder Lesch–Nyhan syndrome (caused by molecular defects within the HPRT gene) (Mansfield et al., 1993). Some data have shown that inheritance of two polymorphic tandem repeats, one being the HPRTB locus (mapped within intron 3 of the HPRT gene), could be used to establish linkage to the disease (Mansfield et al., 1993).

The X chromosome has had an interesting journey in the last two decades in the research fields of forensic and population genetics by providing new (population) data and aiding in the clarification of several issues, namely, in kinship testing. Its particular properties of inheritance (recombination on the female side and haploid state on the male side) have allowed this chromosome a role that cannot be accomplished by the autosomes neither by its counterpart, the Y chromosome. After an initial bloom of publications, several multiplex developments, workshops at international meetings, creation of an X-STR database, the interest in X-chromosomal markers is gradually fading. Analytical and statistical issues may be the major underlined motivations to the lack of interest in addition to a lower demand of X-STR-based identification cases.

Considerable effort has already been put in X-STRs, namely, (i) the generation of allelic and haplotypic frequency databases that include a fair enough number of geographically different located populations; (ii) several in-house multiplexes containing a large number of highly polymorphic markers as well as a sound established commercial kit; and (iii) relevant number of studies addressing and recommending solutions for the main issues surrounding X-STR kinship-based testing. Therefore, this effort should not be lost and move toward the revival of the standing position of X chromosome markers in forensic genetics.

AUTHOR CONTRIBUTIONS

IG drafted the manuscript scheme and, in addition, all authors have made a substantial, direct, and intellectual contribution to the work, and approved the final version for publication.

FUNDING

This work was partially financed by FEDER-Fundo Europeu de Desenvolvimento Regional funds through the COMPETE 2020-Operacional Programme for Competitiveness and Internationalisation (POCI), Portugal 2020, and by Portuguese funds through FCT-Fundação para a Ciência e a Tecnologia/Ministério da Ciência, Tecnologia e Inovação in the framework of the projects "Institute for Research and Innovation in Health Sciences" (POCI-01-0145-FEDER-007274). IG is funded by the FCT Scientific Stimulus program CEECIND/02609/2017. NP and VG are supported by FCT under the program contract provided in Decree-Law no.57/2016 of August 29. SA-S is funded by the FCT doctoral grant SFRH/BD/136284/2018. LG is supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico-CNPq (ref. 306342/2019-7) and Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro-FAPERJ (CNE-2018).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00926/full#supplementary-material

REFERENCES

Amorim, A., and Pereira, L. (2005). Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. *Forensic Sci. Int.* 150, 17–21. doi: 10.1016/j.forsciint.2004.06.018

Amorim, A., and Pinto, N. (2018). Big data in forensic genetics. *Forensic Sci Int Genet.* 37, 102–105. doi: 10.1016/j.fsigen.2018.08.001

Antão-Sousa, S., Amorim, A., Gusmão, L., and Pinto, N. (2019). Mutation in Y STRs: Repeat motif gains vs. losses. *Forensic Science International: Genetics Supplement Series* 7, 240–242. doi: 10.1016/j.fsigss.2019.09.092

Bär, W., Brinkmann, B., Budowle, B., Carracedo, A., Gill, P., Lincoln, P., et al. (1997). DNA recommendations. Further report of the DNA Commission of the ISFH regarding the use of short tandem repeat systems. International Society for Forensic Haemogenetics. *Int J Legal Med* 110, 175–176. doi: 10.1007/s004140050061

Bini, C., Di Nunzio, C., Aneli, S., Sarno, S., Alù, M., Carnevali, E., et al. (2019). Analysis of recombination and mutation events for 12 X-Chr STR loci: A collaborative family study of the Italian Speaking Working Group Ge.F.I. *Forensic Science International: Genetics Supplement Series* 7, 398–400. doi: 10.1016/j.fsigss.2019.10.027

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., and Rolf, B. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *The American Journal of Human Genetics* 62, 1408–1415. doi: 10.1086/301869

Budowle, B., and van Daal, A. (2008). Forensically relevant SNP classes. *Biotechniques* 44, 603–608,610. doi: 10.2144/000112806

Burgos, G., Posada, Y., Florez-Misas, A., Ávila, C., and Ibarra, A. (2019). An update of STR mutation rates from paternity tests analyzed in a 14 year period (2005–2018) at IdentiGEN lab, Universidad de Antioquia, Colombia. *Forensic Science International: Genetics Supplement Series* 7, 530–531. doi: 10.1016/j.fsigss.2019.10.078

Butler, J. M., Coble, M. C., and Vallone, P. M. (2007). STRs vs. SNPs: thoughts on the future of forensic DNA testing. *Forensic sci med pathol* 3, 200–205. doi: 10.1007/s12024-007-0018-1

Carvalho, R., and Pinheiro, M. F. (2011). Study of DXS9895 and DXS7130: Population data from North of Portugal. *J Forensic Leg Med.* 18, 21–22. doi: 10.1016/j.jflm.2010.11.010

Chakraborty, R., Stivers, D. N., Su, B., Zhong, Y., and Budowle, B. (1999). The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. *Electrophoresis* 20, 1682–1696. doi: 10.1002/(SICI)1522-2683(19990101)20:8<1682::AID-ELPS1682<3.0.CO;2-Z

ChrX-Str.org 2.0. (2020). *ChrX-STR.org 2.0*. http://www.chrx-str.org/ (accessed May 15, 2020)

Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P. D., et al. (2014). Origins and functional evolution of Y chromosomes across mammals. *Nature* 508, 488–493. doi: 10.1038/nature13151

Deng, C., Song, F., Li, J., Hou, Y., and Luo, H. (2017). Multiplex PCR for 19 X-chromosomal STRs in Chinese population. *Forensic Science International: Genetics Supplement Series* 6, e24–e26. doi: 10.1016/j.fsigss.2017.09.016

Diegoli, T. M. (2015). Forensic typing of short tandem repeat markers on the X and Y chromosomes. *Forensic Sci. Int. Genet.* 18, 140–151. doi: 10.1016/j.fsigen.2015.03.013

Diegoli, T. M., Rohde, H., Borowski, S., Krawczak, M., Coble, M. D., and Nothnagel, M. (2016). Genetic mapping of 15 human X-chromosomal forensic short tandem repeat (STR) loci by means of multi-core parallelization. *Forensic Sci. Int. Genet.* 25, 39–44. doi: 10.1016/j.fsigen.2016.07.004

Edelmann, J., Kohl, M., Dressler, J., and Hoffmann, A. (2016). X-chromosomal 21-indel marker panel in German and Baltic populations. *Int. J. Legal. Med.* 130, 357–360. doi: 10.1007/s00414-015-1221-3

Edelmann, J., Lessig, R., Willenberg, A., Wildgrube, R., Hering, S., and Szibor, R. (2006). Forensic validation of the X-chromosomal STR-markers GATA165B12, GATA164A09, DXS9908 and DXS7127 in German population. *International Congress Series* 1288, 298–300. doi: 10.1016/j.ics.2005.09.022

Edwards, A., Civitello, A., Hammond, H. A., and Caskey, C. T. (1991). DNA Typing and Genetic Mapping With Trimeric and Tetrameric Tandem Repeats. *Am. J. Hum. Genet.* 49, 746–756.

Edwards, A., Hammond, H. A., Jin, L., Caskey, C. T., and Chakraborty, R. (1992). Genetic Variation at Five Trimeric and Tetrameric Tandem Repeat Loci in Four Human Population Groups. *Genomics* 12, 241–253. doi: 10.1016/0888-7543(92)90371-x

Egeland, T., Pinto, N., and Vigeland, M. D. (2014). A general approach to power calculation for relationship testing. *Forensic Sci. Int. Genet.* 9, 186–190. doi: 10.1016/j.fsigen.2013.05.001

Elakkary, S., Hoffmeister-Ullerich, S., Schulze, C., Seif, E., Sheta, A., Hering, S., et al. (2014). 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 sample. *Forensic Sci Int Genet.* 11, 26–30. doi: 10.1016/j.fsigen.2014.02.007

Fan, G., Ye, Y., Luo, H., and Hou, Y. (2015). Screening of Multi-InDel markers on X-chromosome for forensic purpose. *Forensic Science International: Genetics Supplement Series* 5, e42–e44. doi: 10.1016/j.fsigss.2015.09.017

Fan, G. Y., Ye, Y., and Hou, Y. P. (2016). Detecting a hierarchical genetic population structure via Multi-InDel markers on the X chromosome. *Scientific Reports* 6, 32178.

Ferragut, J. F., Pinto, N., Amorim, A., and Picornell, A. (2019). Improving publication quality and the importance of Post Publication Peer Review: The illustrating example of X chromosome analysis and calculation of forensic parameters. *Forensic Sci Int Genet.* 38, e5–e7. doi: 10.1016/j.fsigen.2018.11.006

Flaquer, A., Fischer, C., and Wienker, T. F. (2009). A new sex-specific genetic map of the human pseudoautosomal regions (PAR1 and PAR2). *Hum Hered.* 68, 192–200. doi: 10.1159/000224639

Freitas, N. S., Resque, R. L., Ribeiro-Rodrigues, E. M., Guerreiro, J. F., Santos, N. P., Ribeiro-Dos-Santos, A., et al. (2010). X-linked insertion/deletion polymorphisms: forensic applications of a 33-markers panel. *Int J Legal Med.* 124, 589–593. doi: 10.1007/s00414-010-0441-9

Fukuta, M., Gaballah, M., Takada, K., Miyazaki, H., Kato, H., Aoki, Y., et al. (2019). Genetic polymorphism of 27 X-chromosomal short tandem repeats in an Egyptian population. *Legal Medicine* 37, 64–66. doi: 10.1016/j.legalmed.2019.01.009

Gao, H., Wang, C., Zhang, R., Wu, H., SunS, Xiao, D., et al. (2019). Application of CPI cutoff value based on parentage testing of duos and trios typed by four autosomal kits. *PLoS ONE* 14:e0225174. doi: 10.1371/journal.pone.0225174

García, M. G., Catanesi, C. I., Penacino, G. A., Gusmão, L., and Pinto, N. (2019). X-chromosome data for 12 STRs: Towards an Argentinian database of forensic haplotype frequencies. *Forensic Science International: Genetics* 41, e8–e13. doi: 10.1016/j.fsigen.2019.04.005

Gettings, K. B., Aponte, R. A., Vallone, P. M., and Butler, J. M. (2015). Current knowledge and future issues. *Forensic Sci Int Genet* 18, 118–130. doi: 10.1016/j.fsigen.2015.06.005

Gill, P., Brenner, C., Brinkmann, B., Budowle, B., Carracedo, A., Jobling, M. A., et al. (2001). DNA Commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs. *Forensic Sci. Int.* 124, 5–10. doi: 10.1016/s0379-0738(01)00498-4

Gjertson, D. W., Brenner, C. H., Baur, M. P., Carracedo, A., Guidet, F., Luque, J. A., et al. (2007). ISFG: recommendations on biostatistics in paternity testing. *Forensic Sci. Int. Genet* 1, 223–231. doi: 10.1016/j.fsigen.2007.06.006

Gomes, C., Magalhães, M., Alves, C., Amorim, A., Pinto, N., and Gusmão, L. (2012). Comparative evaluation of alternative batteries of genetic markers to complement autosomal STRs in kinship investigations: autosomal indels vs. X-chromosome STRs. *Int. J. Legal Med.* 126, 917–921. doi: 10.1007/s00414-012-0768-5

Gomes, I., Brehm, A., Gusmão, L., and Schneider, P. M. (2016). New Sequence Variants Detected at DXS10148, DXS10074 and DXS10134 Loci. *Forensic Sci Int Genet* 20, 112–116. doi: 10.1016/j.fsigen.2015.10.005

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

Gomes, I., Pereira, R., Mayr, W. R., Amorim, A., Carracedo, A., and Gusmão, L. (2009). Evaluation of DXS9902, DXS7132, DXS6809, DXS7133, and DXS7423 in humans and chimpanzees: sequence variation, repeat structure, and nomenclature. *Int. J. Legal Med.* 123, 403–412. doi: 10.1007/s00414-009-0357-4

Gomes, I., Prinz, M., Pereira, R., Bieschke, E., Amorim, A., Carracedo, A., et al. (2008). Sequence variation at three X chromosomal short tandem repeats in Caucasian and African populations. *Forensic Science International: Genetics Supplement Series* 1, 147–149. doi: 10.1016/j.fsigss.2007.10.097

Gunter, C. (2005). Genome Biology: She Moves in Mysterious Ways. *Nature* 434, 279–280. doi: 10.1038/434279a

Gusmão, L., Butler, J. M., Carracedo, A., Gill, P., Kayser, M., Mayr, W. R., et al. (2006). International Society of Forensic Genetics. DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. *Int J Legal Med* 120, 191–200. doi: 10.1007/s00414-005-0026-1

Gusmão, L., Butler, J. M., Linacre, A., Parson, W., Roewer, L., Schneider, P. M., et al. (2017). Revised guidelines for the publication of genetic population data. *Forensic Sci Int Genet.* 30, 160–163. doi: 10.1016/j.fsigen.2017.06.007

Gusmão, L., González-Neira, A., Alves, C., Lareu, M., Costa, S., Amorim, A., et al. (2002). Chimpanzee homologous of human Y specific STRs: A comparative study and a proposal for nomenclature. *Forensic Sci. Int.* 126, 129–136. doi: 10.1016/s0379-0738(02)00046-4

Gusmão, L., Sánchez-Diz, P., Alves, C., Gomes, I., Zarrabeitia, M. T., Abovich, M., et al. (2008). A GEP-ISFG collaborative study on the optimization of an X-STR decaplex: data on 15 Iberian and Latin American populations. *Int. J. Legal Med.* 123, 227–234.

Gusmão, L., Sánchez-Diz, P., Alves, C., Gomes, I., Zarrabeitia, M. T., Abovich, M., et al. (2009). A GEP-ISFG collaborative study on the optimization of an X-STR decaplex: data on 15 Iberian and Latin American populations. *Int. J. Legal Medicine* 123, 227–234. doi: 10.1007/s00414-008-0309-4

Haldane, J. B. S. (1919). The combination of linkage values, and the calculation of distances between the loci of linked factors. *J. Genet.* 8, 299–309.

Hearne, C. M., and Todd, J. A. (1991). Tetranucleotide repeat polymorphism at the HPRT locus. *Nucleic Acids Res.* 19, 5450. doi: 10.1093/nar/19.19.5450

Hedrick, P. W. (1987). Gametic disequilibrium measures: proceed with caution. *Genetics* 117, 331–341.

Hwa, H. L., Chung, W. C., Chen, P. L., Lind, C. L., Lie, H. Y., Yin, H. I., et al. (2018). A 1204-single nucleotide polymorphism and insertion-deletion polymorphism panel for massively parallel sequencing analysis of DNA mixtures. *Forensic Sci Int Genet.* 32, 94–101. doi: 10.1016/j.fsigen.2017.11.002

Hwa, H. L., Wu, M. Y., Lin, C. P., Hsieh, W. H., Yin, H. I., Lee, T. T., et al. (2019). A single nucleotide polymorphism panel for individual identification and ancestry assignment in Caucasians and four East and Southeast Asian populations using a machine learning classifier. *Forensic Sci Med Pathol.* 15, 67–74. doi: 10.1007/s12024-018-0071-y

Jacquard, A. (1974). *The genetic structure of populations; in Biomathematics*, Vol. 5. Berlin: Springer-Verlag, 102–107.

Jäger, A. C., Alvarez, M. L., Davis, C. P., Guzmán, E., Han, Y., Way, L., et al. (2017). Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories. *Forensic Sci Int Genet.* 28, 52–70. doi: 10.1016/j.fsigen.2017.01.011

Jin, B., Su, Q., Luo, H., Li, Y., Wu, J., Yan, J., et al. (2016). Mutational analysis of 33 autosomal short tandem repeat (STR) loci in southwest Chinese Han population based on trio parentage testing. *Forensic Science International: Genetics* 23, 86–90. doi: 10.1016/j.fsigen.2016.03.009

Kauppi, L., Jasin, M., and Keeney, S. (2012). The tricky path to recombining X and Y chromosomes in meiosis. *Ann. N. Y. Acad. Sci.* 1267, 18–23. doi: 10.1111/j.1749-6632.2012.06593.x

Kling, D., Dell'Amico, B., and Tillmar, A. O. (2015a). FamLinkX - implementation of a general model for likelihood computations for X-chromosomal marker data. *Forensic Sci. Int. Genet.* 17, 1–7. doi: 10.1016/j.fsigen.2015.02.007

Kling, D., Tillmar, A., Egeland, T., and Mostad, P. (2015b). A general model for likelihood computations of genetic marker data accounting for linkage, linkage disequilibrium, and mutations. *Int. J. Legal Med.* 129, 943–954. doi: 10.1007/s00414-014-1117-7

Kohn, M., Kehrer-Sawatzki, H., Vogel, W., Graves, J. A., and Hameister, H. (2004). Wide genome comparisons reveal the origins of the human X chromosome. *Trends Genet.* 20, 598–603. doi: 10.1016/j.tig.2004.09.008

Kosambi, D. D. (1944). The estimation of map distances from recombination values. *Ann. Eugen.* 12, 172–175. doi: 10.1111/j.1469-1809.1943.tb02321.x

Koyama, H., Iwasa, M., Tsuchimochi, T., Maeno, Y., Isobe, I., Seko-Nakamura, Y., et al. (2002). Y-STR haplotype data and allele frequency of the DXS10011 locus in a Japanese population sample. *Forensic Sci. Int.* 125, 273–276. doi: 10.1016/s0379-0738(01)00649-1

Krawczak, M. (2007). Kinship testing with X-chromosomal markers: mathematical and statistical issues. *Forensic Sci. Int. Genet.* 1, 111–114. doi: 10.1016/j.fsigen.2007.01.014

Lahn, B. T., and Page, D. C. (1999). Four evolutionary strata on the human X chromosome. *Science (*1999). 286, 964–967. doi: 10.1126/science.286.5441.964

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062

Lazaruk, K., Wallin, J., Holt, C., Nguyen, T., and Walsh, P. S. (2001). Sequence variation in humans and other primates at six short tandem repeat loci used in forensic identity testing. *Forensic Sci. Int.* 119, 1–10. doi: 10.1016/s0379-0738(00)00388-1

Levinson, G., and Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* 4, 203–221. doi: 10.1093/oxfordjournals.molbev.a040442

Li, L., Li, C., Zhang, S., Zhao, S., Liu, Y., and Lin, Y. (2010). Analysis of 14 highly informative SNP markers on X chromosome by TaqMan SNP genotyping assay. *Forensic Sci Int Genet.* 4, e145–e148. doi: 10.1016/j.fsigen.2010.04.004

Li, J., Deng, C., Luo, H., Song, F., and Hou, Y. (2017). *Analyzing an off Forensic Science International: Genetics Supplement Series* 6, e92–e93.

Lin, L., Li, J., Hu, Y., Wang, H., Marah, F. A., Moseray, M., et al. (2020). Genetic characterization of 19 X-STRs in Sierra Leone population from Freetown. *Int. J. Legal Med.* 34, 1659–1661. doi: 10.1007/s00414-019-02243-6

Liu, Q.-L., Wang, J.-Z., Quan, L., Zhao, H., Wu, Y.-D., Huang, X.-L., et al. (2013). Allele and Haplotype Diversity of 26 X-STR Loci in Four Nationality Populations from China. *PLoS ONE* 8:e65570. doi: 10.1371/journal.pone.0065570

Lopes, A. M., Calafell, F., and Amorim, A. (2004). Microsatellite variation and evolutionary history of PCDHX/Y gene pair within the Xq21.3/*Yp*11.2 hominid-specific homology block. *Mol Biol Evol.* 21, 2092–2101. doi: 10.1093/molbev/msh218

Mansfield, E. S., Blasband, A., Kronick, M. N., Wrabetz, L., Kaplan, P., Rappaport, E., et al. (1993). Fluorescent approaches to diagnosis of Lesch-Nyhan syndrome and quantitative analysis of carrier status. *Mol. Cell. Probes* 7, 311–324. doi: 10.1006/mcpr.1993.1045

Matveevsky, S., Kolomiets, O., Bogdanov, A., Hakhverdyan, M., and Bakloushinskaya, I. (2017). Chromosomal Evolution in Mole Voles Ellobius (Cricetidae, Rodentia): Bizarre Sex Chromosomes, Variable Autosomes and Meiosis. *Genes (Basel).* 8, 306. doi: 10.3390/genes8110306

Medina-Acosta, E. (2011). Evidence of partial and weak gametic disequilibrium across clusters of pericentromeric short tandem repeats loci on human X chromosome: proceed with caution in forensic genetics. *Forensic Sci Int Genet.* 5, 545–547. doi: 10.1016/j.fsigen.2009.12.002

Nachman, M. W., and Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics* 156, 297–304.

National Institute of Standards, and Technology [Nist]. (2020). *Standard Reference Database*. Avaliable at: https://strbase.nist.gov/ (accessed May 10, 2020).

Nielsen, J., and Wohlert, M. (1990). Sex chromosome abnormalities found among 34,910 newborn children: results from a 13-year incidence study in Arhus. *Denmark. Birth Defects Original Article Series* 26, 209–223. doi: 10.1007/BF01213097

Nishi, T., Fukui, K., and Iwadate, K. (2020). Genetic Polymorphism Analyses of Three Novel X Chromosomal Short Tandem Repeat Loci in the Xp22.3 Region. *Legal Medicine* 45, 101709. doi: 10.1016/j.legalmed.2020.101709

Nothnagel, M., Szibor, R., Vollrath, O., Augustin, C., Edelmann, J., Geppert, M., et al. (2012). Collaborative genetic mapping of 12 forensic short tandem repeat (STR) loci on the human X chromosome. *Forensic Sci. Int. Genet.* 6, 778–784. doi: 10.1016/j.fsigen.2012.02.015

Ohta, T., and Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res.* 22, 201–204. doi: 10.1073/pnas.75.6.2868

Oki, T., Hayashi, T., Ota, M., and Asamura, H. (2012). Development of multiplex assay with 16 SNPs on X chromosome for degraded samples. *Leg Med (Tokyo)* 14, 11–16. doi: 10.1016/j.legalmed.2011.10.001

Olaisen, B., Bär, W., Brinkmann, B., Budowle, B., Carracedo, A., Gill, P., et al. (1998). DNA recommendations 1997 of the International Society for Forensic Genetics. *Vox Sang* 74, 61–63.

Otto, S., Pannell, J. R., Peichel, C. L., Ashman, T. L., Charlesworth, D., Chippindale, A. K., et al. (2011). About PAR: the distinct evolutionary dynamics of the pseudoautosomal region. *Trends Genet.* 27, 358–367. doi: 10.1016/j.tig.2011.05.001

Parson, W., and Roewer, L. (2010). Publication of Population Data of Linearly Inherited DNA Markers in the International Journal of Legal Medicine. *Int J Legal Med.* 124, 505–509. doi: 10.1007/s00414-010-0492-y

Pereira, R., Alves, C., Aler, M., Amorim, A., Arévalo, C., Betancor, E., et al. (2018). A GHEP-ISFG collaborative study on the genetic variation of 38 autosomal indels for human identification in different continental populations. *Forensic Sci Int Genet.* 32, 18–25. doi: 10.1016/j.fsigen.2017.09.012

Pereira, R., Gomes, I., Amorim, A., and Gusmão, L. (2007). Genetic diversity of 10 X chromosome STRs in northern Portugal. *Int. J. Legal Med.* 121, 192–197. doi: 10.1007/s00414-006-0144-4

Pereira, R., Pereira, V., Gomes, I., Tomas, C., Morling, N., Amorim, A., et al. (2012). A method for the analysis of 32 X chromosome insertion deletion polymorphisms in a single PCR. *Int J Legal Med.* 126, 97–105. doi: 10.1007/s00414-011-0593-2

Pereira, R., Phillips, C., Alves, C., Amorim, A., Carracedo, A., and Gusmão, L. (2009). A new multiplex for human identification using insertion/deletion polymorphisms. *Electrophoresis.* 30, 3682–3690. doi: 10.1002/elps.200900274

Pereira, V., and Gusmão, L. (2016). "Types of Genomes, Sequences and Genetic Markers (Repeats, SNPs, Indels, Haplotypes)," in *Handbook Of Forensic Genetics: Biodiversity And Heredity In Civil And Criminal Investigation*, Vol. 2, (Singapore: World Scientific), 163. doi: 10.1142/9781786340788_0009

Petkovski, E., Keyser-Tracqui, C., Hienne, R., and Ludes, B. (2005). SNPs and MALDI-TOF MS: tools for DNA typing in forensic paternity testing and anthropology. *J Forensic Sci.* 50, 535–541.

Phillips, C., Ballard, D., Gill, P., Syndercombe Court, D., Carracedo, A., and Lareu, M. V. (2012). The recombination landscape around forensic STRs: Accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data. *Forensic Science International: Genetics* 6, 354–365. doi: 10.1016/j.fsigen.2011.07.012

Phillips, C., Devesse, L., Ballard, D., van Weert, L., de la Puente, M., and Melis, S. (2018). Global patterns of STR sequence variation: Sequencing the CEPH human genome diversity panel for 58 forensic STRs using the Illumina ForenSeq DNA Signature Prep Kit. *Electrophoresis* 39, 2708–2724. doi: 10.1002/elps.201800117

Pinto, N., Gusmão, L., and Amorim, A. (2011a). X-chromosome markers in kinship testing: a generalisation of the IBD approach identifying situations where their contribution is crucial. *Forensic Sci. Int. Genet.* 5, 27–32. doi: 10.1016/j.fsigen.2010.01.011

Pinto, N., Gusmão, L., Silva, P. V., and Amorim, A. (2011b). Estimating coancestry from genotypes using a linear regression method. *Forensic Science International: Genetics Supplement Series* 3, e373–e374. doi: 10.1016/j.fsigss.2011.09.048

Pinto, N., Gusmão, L., Egeland, L. T., and Amorim, A. (2013a). Paternity exclusion power: comparative behaviour of autosomal and X-chromosomal markers in standard and deficient cases with inbreeding. *Forensic Sci. Int. Genet.* 7, 290–295. doi: 10.1016/j.fsigen.2012.12.002

Pinto, N., Gusmão, L., Egeland, T., and Amorim, A. (2013b). Estimating relatedness with no prior specification of any genealogy: The role of the X-chromosome. *Forensic Science International: Genetics Supplement Series* 4, e252–e253. doi: 10.1016/j.fsigss.2013.10.129

Pinto, N., Silva, P. V., and Amorim, A. (2010). General Derivation of the Sets of Pedigrees with the Same Kinship Coefficients. *Hum. Hered.* 70, 194–204. doi: 10.1159/000316390

Pinto, N., Silva, P. V., and Amorim, A. (2012). A general method to assess the utility of the X-chromosomal markers in kinship testing. *Forensic Science International: Genetics* 6, 198–207. doi: 10.1016/j.fsigen.2011.04.014

Prieto-Fernández, E., Baeta, M., Núñez, C., Zarrabeitia, M. T., and Herrera, R. J. (2016). Development of a new highly efficient 17 X-STR multiplex for forensic purposes. *Electrophoresis.* 37, 1651–1658. doi: 10.1002/elps.201500546

Primmer, C. R., Saino, N., Møller, A. P., and Ellegren, H. (1996). Directional evolution in germline microsatellite mutations. *Nat. Genet.* 13, 391. doi: 10.1038/ng0896-391

Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., McLay, K., Muzny, D., et al. (2005). The DNA sequence of the human X chromosome. *Nature* 434, 325–337. doi: 10.1038/nature03440

Schaffner, S. F. (2004). The X chromosome in population genetics. *Nat. Rev. Genet.* 5, 43–51. doi: 10.1038/nrg1247

Schlötterer, C. (2000). Evolutionary dynamics of microsatellite DNA. *Chromosoma* 109, 365–371. doi: 10.1007/s004120000089

Schlötterer, C., and Tautz, D. (1992). Slippage synthesis of simple sequence DNA. *Nucleic Acids Res* 20, 211–215. doi: 10.1093/nar/20.2.211

Shimmin, L. C., Chang, B. H., and Li, W. H. (1993). Male-driven evolution of DNA sequences. *Nature* 362, 745–747. doi: 10.1038/362745a0

Sleddens, H. F., Oostra, B. A., Brinkmann, A. O., and Trapman, J. (1992). Trinucleotide repeat polymorphism in the androgen receptor gene (AR). *Nucleic Acids Res.* 20, 1427. doi: 10.1093/nar/20.6.1427-a

Stepanov, V., Vagaitseva, K., Kharkov, V., Cherednichenko, A., Bocharova, A., Berezina, G., et al. (2016). Forensic and population genetic characteristics of 62 X chromosome SNPs revealed by multiplex PCR and MALDI-TOF mass spectrometry genotyping in 4 North Eurasian populations. *Leg Med (Tokyo).* 18, 66–71. doi: 10.1016/j.legalmed.2015.12.008

Strand, M., Prolla, T. A., Liskay, R. M., and Petes, T. D. (1993). Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* 365, 274. doi: 10.1038/365274a0

STRidER. (2020). *(STRs for Identity ENFSI Reference Database).* Avaliable at: https://strider.online/ (accessed May 10, 2020).

Szibor, R. (2007). X-chromosomal markers: past, present and future. *Forensic Sci Int Genet* 1, 93–99. doi: 10.1016/j.fsigen.2007.03.003

Szibor, R., Edelmann, J., Hering, S., Gomes, I., and Gusmão, L. (2009). Nomenclature Discrepancies in the HPRTB Short Tandem Repeat. *Int. J. Legal Med.* 123, 185–186. doi: 10.1007/s00414-008-0314-7

Szibor, R., Hering, S., and Edelmann, J. (2005). The HumARA genotype is linked to spinal and bulbar muscular dystrophy and some further disease risks and should no longer be used as a DNA marker for forensic purposes. *Int. J. Legal Med.* 119, 179–180. doi: 10.1007/s00414-005-0525-0

Szibor, R., Hering, S., and Edelmann, J. (2006). A new Web site compiling forensic chromosome X research is now online. *Int J Legal Med.* 120, 252–254. doi: 10.1007/s00414-005-0029-y

Szibor, R., Krawczak, M., Hering, S., Edelmann, J., Kuhlisch, E., and Krause, D. (2003). Use of X-linked markers for forensic purposes. *Int. J. Legal Med.* 117, 67–74. doi: 10.1007/s00414-002-0352-5

Tao, R., Zhang, J., Sheng, X., Zhang, J., Yang, Z., Chen, C., et al. (2019). Development and validation of a multiplex insertion/deletion marker panel, SifaInDel 45plex system. *Forensic Science International: Genetics* 41, 128–136. doi: 10.1016/j.fsigen.2019.04.008

The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 *million SNPs. Nature* 449, 851–861. doi: 10.1038/nature06258

The International Society for Forensic Genetics [ISFG]. (2020). Avaliable at: www.isfg.org (accessed May 10, 2020).

Tillmar, A. O., Egeland, T., Lindblom, B., Holmlund, G., and Mostad, P. (2011). Using X-chromosomal markers in relationship testing: calculation of likelihood ratios taking both linkage and linkage disequilibrium into account. *Forensic Sci. Int. Genet.* 5, 506–511. doi: 10.1016/j.fsigen.2010.11.004

Tillmar, A. O., Kling, D., Butler, J. M., Parson, W., Prinz, M., Schneider, P. M., et al. (2017). DNA Commission of the International Society for Forensic Genetics (ISFG): Guidelines on the use of X-STRs in kinship analysis. *Forensic Sci Int Genet.* 29, 269–275. doi: 10.1016/j.fsigen.2017.05.005

Tomas, C., Pereira, V., and Morling, N. (2012). Analysis of 12 X-STRs in Greenlanders, Danes and Somalis using Argus X-12. *Int J Legal Med.* 126, 121–128. doi: 10.1007/s00414-011-0609-y

Tomas, C., Sanchez, J. J., Castro, J. A., Børsting, C., and Morling, N. (2010). Forensic usefulness of a 25 X-chromosome single-nucleotide polymorphism marker set. *Transfusion.* 50, 2258–2265. doi: 10.1111/j.1537-2995.2010.02696.x

Watanabe, G., Umetsu, K., Yuasa, I., and Suzuki, T. (2000). DXS10011: a hypervariable tetranucleotide STR polymorphism on the X chromosome. *Int J Legal Med.* 113, 249–250. doi: 10.1007/s004149900096

Weir, B., Anderson, A., and Hepler, A. (2006). Genetic relatedness analysis: modern data and new challenges. *Nat. Rev. Genet.* 7, 771–780. doi: 10.1038/nrg1960

Xu, X., Peng, M., and Fang, Z. (2000). The direction of microsatellite mutations is dependent upon allele length. *Nat. Genet.* 24, 396. doi: 10.1038/74238

YHRD. (2020). *Y-STR Haplotype Reference Database (*YHRD). Avaliable at: https://yhrd.org/YHRD (accessed May 10, 2020).

Zarrabeitia, M. T., Mijares, V., and Riancho, J. A. (2007). Forensic efficiency of microsatellites and single nucleotide polymorphisms on the X chromosome. *Int J Legal Med.* 121, 433–437. doi: 10.1007/s00414-007-0169-3

Zaumsegel, D., Rothschild, M. A., and Schneider, P. M. (2013). A 21 marker insertion deletion polymorphism panel to study biogeographic ancestry. *Forensic Sci Int Genet.* 7, 305–312. doi: 10.1016/j.fsigen.2012.12.007

Zhang, S., Bian, Y., Chen, A., Zheng, H., Gao, Y., Hou, Y., et al. (2017a). Developmental validation of a custom panel including 273 SNPs for forensic application using Ion Torrent PGM. *Forensic Sci Int Genet.* 27, 50–57. doi: 10.1016/j.fsigen.2016.12.003

Zhang, S., Lin, Y., Bian, Y., and Li, C. (2017b). Parallel sequencing of 60 X-chromosome genetic markers including STRs, SNPs and InDels. *Forensic Science International: Genetics Supplement Series* 6, e317–e319. doi: 10.1016/j.fsigss.2017.09.127

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A

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*, 926. doi: 10.3389/fgene.2020.00926

Supplementary Table 1. X chromosome specific STR polymorphisms described in the literature till date. The reference given is not the original report, but the most recent usage of the marker. Literature search was performed using Scopus database (https://www.scopus.com/search/form.uri?display=basic), PubMed Central (PMC), U.S. National Institutes of Health's National Library of Medicine. Proceeding reports were found using FSI:Genetics Journal's website directly.

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	-

Locus	Reference	Aliases/ comments
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

Locus	Reference	Aliases/ comments
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	-

Locus	Reference	Aliases/ comments
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	-

Locus	Reference	Aliases/ comments
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.	-

Locus	Reference	Aliases/ comments
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	-

Supplementary Table 2: Compilation of X-STR population data available.

Accessible online at https://doi.org/10.3389/fgene.2020.00926.

Supplementary Table 3: Compilation of marker mutation rates per familial configuration and respective confidence interval (95%) reported in the literature.

	Mother-son duos				Mother-daughter duos				Father-daughter duos				Trios with daughter				
Markers	Meioses	Mutations	Mutation Rate	CI (95%)	Meioses	Mutations	Mutation Rate	CI (95%)	Meioses	Mutations	Mutation Rate	CI (95%)	Meioses	Mutations	Mutation Rate	CI (95%)	References
DXS101	737	0	0,000	0,000-0,005	40	0	0,000	0,000-0,088	33	0	0,000	0,000-0,106	1021	1	0,001	0,000-0,005	[1, 4, 7, 9, 12, 16, 18]
DXS7424	940	0	0,000	0,000-0,004	335	0	0,000	0,000-0,011	186	0	0,000	0,000-0,020	975	2	0,002	0,000-0,007	[2, 7, 12, 16, 18, 20]
DXS6803	1034	0	0,000	0,000-0,004	335	0	0,000	0,000-0,011	165	0	0,000	0,000-0,022	847	3	0,004	0,001-0,010	[3, 10, 18, 20]
DXS9895	-	-	-	-	-	-	-	-	-	-	-	-	78	0	0,000	0,000-0,046	[3]
DXS8377	77	1	0,013	0,000-0,070	-	-	-	-	12	0	0,000	0,000-0,265	134	1	0,007	0,000-0,041	[4, 8, 9, 12]
HPRTB	904	1	0,001	0,000-0,006	335	1	0,003	0,000-0,017	329	3	0,009	0,002-0,026	2946	7	0,002	0,001-0,005	[4, 6, 7, 8, 9, 11, 12, 14, 15, 18, 20, 21]
DXS7132	1332	1	0,001	0,000-0,004	335	1	0,003	0,000-0,017	329	2	0,006	0,001-0,022	3577	32	0,009	0,006-0,013	[5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16, 17, 18, 19, 20, 21]
DXS7133	177	0	0,000	0,000-0,021	-	-	-	-	21	0	0,000	0,000-0,161	474	0	0,000	0,000-0,008	[5, 7, 12, 13, 16]
GATA172D05	820	0	0,000	0,000-0,004	335	0	0,000	0,000-0,011	176	0	0,000	0,000-0,021	676	2	0,003	0,000-0,011	[5, 7, 8, 9, 17, 18, 20]
DXS8378	891	2	0,002	0,000-0,008	335	0	0,000	0,000-0,011	329	0	0,000	0,000-0,011	2933	6	0,002	0,001-0,004	[6, 7, 8, 9, 11, 12, 14, 15, 17, 18, 19, 20, 21]
DXS7423	618	0	0,000	0,000-0,006	40	0	0,000	0,000-0,088	175	0	0,000	0,000-0,021	2875	4	0,001	0,000-0,004	[5, 6, 7, 9, 11, 12, 14, 15, 17, 18, 19, 21]

	Mother-son duos				Mother-daughter duos				Father-daughter duos				Trios with daughter				
Markers	Meioses	Mutations	Mutation Rate	CI (95%)	Meioses	Mutations	Mutation Rate	CI (95%)	Meioses	Mutations	Mutation Rate	CI (95%)	Meioses	Mutations	Mutation Rate	CI (95%)	References
DXS6809	498	2	0,004	0,000-0,014	-	-	-	-	11	0	0,000	0,000-0,285	715	4	0,006	0,002-0,014	[7, 9, 10, 16, 17]
DXS6789	1261	0	0,000	0,000-0,003	335	0	0,000	0,000-0,011	176	0	0,000	0,000-0,021	1240	5	0,004	0,001-0,009	[7, 8, 9, 10, 16, 17, 18, 19, 20]
GATA31E08	947	0	0,000	0,000-0,004	335	0	0,000	0,000-0,011	175	0	0,000	0,000-0,021	895	1	0,001	0,000-0,006	[7, 8, 16, 17, 18, 20]
DXS6807	280	0	0,000	0,000-0,013	295	0	0,000	0,000-0,012	175	0	0,000	0,000-0,021	98	0	0,000	0,000-0,037	[7, 12, 20]
DXS9902	777	0	0,000	0,000-0,005	335	0	0,000	0,000-0,011	165	0	0,000	0,000-0,022	545	5	0,009	0,003-0,021	[8, 17, 18, 20]
DXS6810	273	0	0,000	0,000-0,013	295	0	0,000	0,000-0,012	154	0	0,000	0,000-0,024	108	1	0,009	0,000-0,051	[8, 20]
DXS981	468	0	0,000	0,000-0,008	-	-	-	-	-	-	-	-	771	0	0,000	0,000-0,005	[4, 8, 10, 13, 16]
DXS6793	-	-	-	-	-	-	-	-	-	-	-	-	50	0	0,000	0,000-0,071	[8]
DXS6801	170	0	0,000	0,000-0,021	-	-	-	-	-	-	-	-	360	0	0,000	0,000-0,010	[8, 16]
DXS9898	234	0	0,000	0,000-0,016	-	-	-	-	12	0	0,000	0,000-0,265	381	0	0,000	0,000-0,010	[9, 12, 16, 17]
DXS10135	64	1	0,016	0,000-0,084	-	-	-	-	46	1	0,022	0,001-0,115	2297	45	0,020	0,014-0,026	[11, 14, 15, 19, 21]
DXS10074	248	0	0,000	0,000-0,015	-	-	-	-	46	0	0,000	0,000-0,077	2627	23	0,009	0,006-0,013	[11, 14, 15, 16, 17, 19, 21]

	Mother-son duos				Mother-daughter duos				Father-daughter duos				Trios with daughter				
Markers	Meioses	Mutations	Mutation Rate	CI (95%)	Meioses	Mutations	Mutation Rate	CI (95%)	Meioses	Mutations	Mutation Rate	CI (95%)	Meioses	Mutations	Mutation Rate	CI (95%)	References
DXS10101	64	0	0,000	0,000-0,056	-	-	-	-	46	0	0,000	0,000-0,077	2297	14	0,006	0,003-0,010	[11, 14, 15, 19, 21]
DXS10134	351	0	0,000	0,000-0,010	295	0	0,000	0,000-0,012	200	1	0,005	0,000-0,028	2375	21	0,009	0,005-0,013	[11, 14, 15, 17, 19, 20, 21]
DXS6800	450	0	0,000	0,000-0,008	295	0	0,000	0,000-0,012	165	0	0,000	0,000-0,022	452	0	0,000	0,000-0,008	[12, 13, 16, 20]
DXS6797	-	-	-	-	-	-	-	-	-	-	-	-	84	0	0,000	0,000-0,043	[13]
GATA165B12	933	0	0,000	0,000-0,004	335	0	0,000	0,000-0,011	165	0	0,000	0,000-0,022	869	0	0,000	0,000-0,004	[13, 16, 18, 20]
DXS10148	64	0	0,000	0,000-0,056	-	-	-	-	46	0	0,000	0,000-0,077	2069	19	0,009	0,006-0,014	[14, 19, 21]
DXS10079	234	1	0,004	0,000-0,024	-	-	-	-	46	0	0,000	0,000-0,077	2379	23	0,010	0,006-0,014	[14, 16, 19, 21]
DXS10103	64	0	0,000	0,000-0,056	-	-	-	-	46	0	0,000	0,000-0,077	2069	9	0,004	0,002-0,008	[14, 19, 21]
DXS10146	64	0	0,000	0,000-0,056	-	-	-	-	46	1	0,022	0,001-0,115	2069	13	0,006	0,003-0,011	[14, 19, 21]
DXS10075	170	0	0,000	0,000-0,021	-	-	-	-	-	-	-	-	310	3	0,010	0,002-0,028	[16]
DXS10147	504	0	0,000	0,000-0,007	40	0	0,000	0,000-0,088	11	0	0,000	0,000-0,285	437	0	0,000	0,000-0,008	[17, 18]
DXS6795	763	0	0,000	0,000-0,005	335	0	0,000	0,000-0,011	165	0	0,000	0,000-0,022	475	0	0,000	0,000-0,008	[18, 19]

	Mother-son duos				Mother-daughter duos				Father-daughter duos				Trios with daughter				
Markers	Meioses	Mutations	Mutation Rate	CI (95%)	Meioses	Mutations	Mutation Rate	CI (95%)	Meioses	Mutations	Mutation Rate	CI (95%)	Meioses	Mutations	Mutation Rate	CI (95%)	References
DXS7130	490	0	0,000	0,000-0,008	40	0	0,000	0,000-0,088	11	0	0,000	0,000-0,285	417	0	0,000	0,000-0,009	[18]
DXS10159	273	0	0,000	0,000-0,013	295	0	0,000	0,000-0,012	154	1	0,00649	0,000-0,036	58	0	0,000	0,000-0,062	[20]

NOTE: Mutations rates were calculated based on the following references: [1] Edelmann *et al.*, 2001; [2] Edelmann *et al.*, 2002; [3] Huang *et al.*, 2003; [4] Athanasiadou *et al.*, 2003; [5] Turrina *et al.*, 2004; [6] Zalan *et al.*, 2007; [7] Turrina *et al.*, 2007; [8] Tariq *et al.*, 2008; [9] Pico *et al.*, 2008; [10] Liu *et al.*, 2008; [11] Becker *et al.*, 2008; [12] Poetsch *et al.*, 2009; [13] Nadeem *et al.*, 2009; [14] Pamjav *et al.*, 2011; [15] Teztaff *et al.*, 2012; [16] Liu *et al.*, 2012; [17] Nishi *et al.*, 2013; [18] Diegoli *et al.*, 2014; [19] García *et al.*, 2018; [20] Chen *et al.*, 2019; [21] Pinto *et al.*, 2020

Athanasiadou, D., *et al.*, (2003). Development of a quadruplex PCR system for the genetic analysis of X-chromosomal STR loci. In *International Congress Series* (Vol. 1239, pp. 311-314). Elsevier.

Becker, D., *et al.*, (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.

Chen, M., *et al.*, (2020). 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*, *134*(1), 163-168.

Diegoli, T. M., *et al.*, (2014). Mutation rates of 15 X chromosomal short tandem repeat markers. *International journal of legal medicine*, *128*(4), 579-587.

Edelmann, J., & Szibor, R. (2001). DXS101: a highly polymorphic X-linked STR. *International journal of legal medicine*, *114*(4-5), 301-304.

Edelmann, J., *et al.*, (2002). Validation of the STR DXS7424 and the linkage situation on the X-chromosome. *Forensic science international*, *125*(2-3), 217-222.

García, M. G., *et al.*, (2019). X-chromosome data for 12 STRs: Towards an Argentinian database of forensic haplotype frequencies. *Forensic Science International: Genetics*, *41*, e8-e13.

Huang, D., *et al.*, (2003). Development of the X-linked tetrameric microsatellite markers HumDXS6803 and HumDXS9895 for forensic purpose. *Forensic science international*, *133*(3), 246-249.

Liu, Q. L., *et al.*, (2008). Development of a five ChX STRs loci typing system. *International journal of legal medicine*, *122*(3), 261-265.

Liu, Q. L., *et al.*, (2012). Development of multiplex PCR system with 15 X-STR loci and genetic analysis in three nationality populations from C hina. *Electrophoresis*, *33*(8), 1299-1305.

Nadeem, A., *et al.*, (2009). Development of pentaplex PCR and genetic analysis of X chromosomal STRs in Punjabi population of Pakistan. *Molecular biology reports*, *36*(7), 1671-1675.

Nishi, T., *et al.*, (2013). Application of a novel multiplex polymerase chain reaction system for 12 X-chromosomal short tandem repeats to a Japanese population study. *Legal Medicine*, *15*(1), 43-46.

Pamjav, H., *et al.*, (2012). X chromosomal recombination study in three-generation families in Hungary. *Forensic Science International: Genetics*, *6*(3), e95-e96.

Pico, A., *et al.*, (2008). Genetic profile characterization and segregation analysis of 10 X-STRs in a sample from Santander, Colombia. *International journal of legal medicine*, *122*(4), 347-351.

Pinto, N., *et al.*, (2020). Paternal and maternal mutations in X-STRs: A GHEP-ISFG collaborative study. *Forensic Science International: Genetics*, *46*, 102258.

Poetsch, M., *et al.*, (2009). Allele frequencies of 11 X-chromosomal loci in a population sample from Ghana. *International journal of legal medicine*, *123*(1), 81.

Tariq, M. A., *et al.*, (2008). Allele frequency distribution of 13 X-chromosomal STR loci in Pakistani population. *International journal of legal medicine*, *122*(6), 525-528.

Tetzlaff, S., Wegener, R., & Lindner, I. (2012). Population genetic investigation of eight X-chromosomal short tandem repeat loci from a northeast German sample. *Forensic Science International: Genetics*, *6*(6), e155-e156.

Turrina, S., & De Leo, D. (2004, April). Population genetic comparisons of three X-chromosomal STRs (DXS7132, DXS7133 and GATA172D05) in North and South Italy. In *International Congress Series* (Vol. 1261, pp. 302-304). Elsevier.

Turrina, S., *et al.*, (2007). Development and forensic validation of a new multiplex PCR assay with 12 X-chromosomal short tandem repeats. *Forensic Science International: Genetics*, *1*(2), 201-204.

Zalán, A., *et al.*, (2007). Hungarian population data of four X-linked markers: DXS8378, DXS7132, HPRTB, and DXS7423. *International journal of legal medicine*, *121*(1), 74-77.

2.1.2. Mutation in Y STRs: Repeat motif gains vs. losses

Antão-Sousa, S., Amorim, A., Gusmão, L., & Pinto, N. (2019). Mutation in Y STRs: Repeat motif gains vs. losses. *Forensic Science International: Genetics Supplement Series*, *7*(1), 240-242. (https://doi.org/10.1016/j.fsigss.2019.09.092)

Forensic Science International: Genetics Supplement Series 7 (2019) 240–242

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

journal homepage: www.elsevier.com/locate/fsigss

Mutation in Y STRs: Repeat motif gains vs. losses

S. Antão-Sousa[a,b,c,d,*], A. Amorim[a,b,c], L. Gusmão[d], N. Pinto[b,c,e]

[a] Faculty of Sciences of the University of Porto (FCUP), Portugal
[b] Institute of Pathology and Molecular Immunology from University of Porto (IPATIMUP), Portugal
[c] Institute of Innovation and Research in Health (i3S), University of Porto, Portugal
[d] DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Brazil
[e] Center of Mathematics of the University of Porto (CMUP), Portugal

ARTICLE INFO

Keywords:
Y-STRs
Mutation's direction
Repeat motif
S
ingle-step mutation
Multistep mutation

ABSTRACT

Y chromosome specific short tandem repeats (Y-STRs) are widely used in population genetics and forensics. Since these markers do not recombine, mutation is the only source of diversity. The primary mutational mechanism leading to length changes in STRs is thought to be polymerase template slippage, and the most common change is the gain or the loss of one repeat motif. In this work, we aim to study 19 Y-STR alleles' contraction and expansion. Alleles were grouped into tertiles: short (1st tertile), intermediate (2nd tertile) and long alleles (3rd tertile). Significant differences between repeat gains and losses were found at four markers - DYS19, DYS439 for intermediate alleles, and DYS570 and DYS626 for long alleles. When the average number is computed for the pooled loci, for short alleles, the number of repeat motif gains is higher than of repeat losses, and the opposite happens for long alleles. For intermediate alleles, the proportion between the number of repeat gains and losses is close to one. Generally, the rate of expansion decreases from the first tertile to the third, and conversely, the rate of contraction increases from the first tertile to the third. The pooled loci tertiles' mutation rate increases from short to long alleles. Our results demonstrate that the mutation direction and rate depend on alleles' length. The longer the allele the greater the mutation and contraction rates.

1. Introduction

Short tandem repeat (STR) loci, also known as microsatellites, are tandemly arrayed repeats of DNA fragments of 1–6 base pairs per repeat motif. STRs have been widely used in population and forensic genetics research. The increasing application of these markers requires the study of mutation rates and mechanisms, as this knowledge is an essential prerequisite for the accurate interpretation of experimental data.

The non-recombining region of the Y chromosome acts as a unique tool for forensic investigations as it is inherited through the patrilineal line without recombination. Thus, this haploid system allows the unambiguous detection of any mutation through the identification of both paternal and filial alleles. Other genetic systems, such as the X-chromosomal (haplodiploid) or autosomal (diploid), always entail uncertainty concerning the identification of either the parental or the mutated allele or both [1–3].

The primary mutational mechanism leading to changes in microsatellite length is though to be polymerase template slippage [4,5] and most of the observed changes in length are due to the gain or loss of a single repeat [6,7]. The stepwise mutation model (SMM) [8] has been widely used to model STR evolution [9–12]. According to this model, the length of a microsatellite varies at a fixed rate independent of length and with the same probability of expansion and contraction. Xu et al. [7] have indeed observed that the rate of expansion mutations is independent of allele length, but several studies [e.g. 7,13,14] report the increase of the overall mutation rate with allele length.

2. Material and methods

We have analyzed 104,083 allele transfers between father-son duos at 19 Y-STR loci.

For each marker, alleles were clustered in tertiles (short, intermediate, and long) considering the number of repeats. The proportion of gains and losses of repeats and the expansion and contraction rates were computed for each tertile for each marker (see Table 1). Overall tertile number of expansions/contractions and mutation rates were also computed.

* Corresponding author at: Faculty of Sciences of the University of Porto (FCUP), Portugal.
E-mail address: sofias@ipatimup.pt (S. Antão-Sousa).

https://doi.org/10.1016/j.fsigss.2019.09.092
Received 6 September 2019; Accepted 26 September 2019
Available online 27 September 2019
1875-1768/

Table 1
Allelic transfers in father-son duos; repeat number gains and losses, and rates of expansion/contraction, per marker and per allelic length tertile.

		1st tercile		2nd tercile		3rd tercile			1st tercile		2nd tercile		3rd tercile	
Marker	N	No. of repeat gains	No. of repeat losses	No. of repeat gains	No. of repeat losses	No. of repeat gains	No. of repeat losses	TOTAL	Rate of expansion	Rate of contraction	Rate of expansion	Rate of contraction	Rate of expansion	Rate of contraction
DYS19	9,074	2	0	**13**[a]	**2**[a]	1	5	23	1.00	0.00	0.87	0.13	0.17	0.83
DYS391	9,108	0	0	13	9	0	2	24			0.59	0.41	0.00	1.00
DYS389I	8,084	0	0	10	10	0	2	22			0.50	0.50	0.00	1.00
DYS460	2,002	1	0	3	4	0	2	10	1.00	0.00	0.43	0.57	0.00	1.00
DYS533	1,085	0	0	1	1	0	1	3			0.50	0.50	0.00	1.00
GATA H4	7,128	0	0	4	7	0	2	13			0.36	0.64	0.00	1.00
DYS439	7,355	2	0	**10**[a]	**1**[a]	6	14	33	1.00	0.00	0.91	0.09	0.30	0.70
DYS549	104	0	0	0	1	0	0	1			0.00	1.00		
DYS393	7,861	0	0	5	2	1	3	11			0.71	0.29	0.25	0.75
GATA A10	874	1	1	0	2	0	0	4	0.50	0.50	0.00	1.00		
DYS456	6,036	1	0	11	8	0	2	22	1.00	0.00	0.58	0.42	0.00	1.00
DYS437	7,333	2	0	2	4	1	1	10	1.00	0.00	0.33	0.67	0.50	0.50
DYS461	873	0	0	0	0	0	0	0						
DYS385	15,341	8	2	11	5	0	1	27	0.80	0.20	0.69	0.31	0.00	1.00
DYS458	6,043	3	3	16	15	6	2	45	0.50	0.50	0.52	0.48	0.75	0.25
DYS570	4,293	2	0	16	20	**0**[a]	**5**[a]	43	1.00	0.00	0.44	0.56	0.00	1.00
DYS576	4,194	1	0	31	24	3	6	65	1.00	0.00	0.56	0.44	0.33	0.67
DYS626	3,138	0	2	9	11	**2**[a]	**10**[a]	34	0.00	1.00	0.45	0.55	0.17	0.83
DYS627	4,157	0	0	21	27	6	12	66			0.44	0.56	0.33	0.67

[a] Significant differences between gains and losses in the tertile.

3. Results

Statistically significant differences between gains and losses were observed in the first tertile and in the third only at DYS570 and DYS626 markers, while in the second tertile differences were verified at DYS19 and DYS439 loci.

Xu et al. [4] observed an increase in the rate of contraction with the increase in allele length, the rate of expansion remaining constant. Our observations support these findings for the rate of contraction, with the exception of markers DYS437, DYS458 and DYS626. However, our data do not support the conclusions concerning the rate of expansion, as it decreases with allele length for 16 out of the 19 markers (excluding DYS437, DYS458 and DYS626) (see Table 1).

Several studies [7,13,14] report that overall mutation rate increase with allele length. Indeed, when the overall mutation rate per tertile was computed, an increase from short (0.00150) to intermediate (0.00531) and long alleles (0.01468) was found.

4. Discussion and conclusions

Only DYS570 and DYS626 markers have shown statistically different number of losses over gains (in the third tertile). Markers DYS19 and DYS439 have shown significantly different number of gains over losses (in the second tertile). No other markers have shown significant differences between gains and losses. When the average number of repeat motif gains and losses for the tertile-pooled loci is computed, for short alleles, the average number of repeat motif gains is higher than the average number of repeat losses ($p = 0.000072$), and the opposite happens for long alleles ($p = 0.000144$). For intermediate alleles, the proportion between the number of repeat gains and losses is close to one ($p = 0.624137$).

Generally, the rate of expansion decreases from the first tertile (0.80) to the second (0.49) and from the second to the third (0.18). Conversely, the rate of contraction increases from the first tertile (0.20) to the second (0.51) and from the second to the third (0.83).

Our results indicate that the proportion of repeat gains and losses varies with allele length. Namely, shorter alleles tend to expand, longer alleles tend to contract, and intermediate alleles have approximately equal tendency to contract or expand.

Furthermore, mutation rate increases as the alleles' length increases. The highest mutation rate was found for long alleles, followed by intermediate alleles and, finally, short alleles.

Our results support the observations that the mutation direction and rate depend on alleles' length. Long alleles have higher mutation and contraction rates and short alleles have lower mutation and higher expansion rates.

Declaration of Competing Interest

The authors declare no conflict of interests.

Acknowledgements

This work was partially financed by FEDER—Fundo Europeu de Desenvolvimento Regional - funds through the COMPETE 2020—Operacional Programme for Competitiveness and Internationalisation (POCI), Portugal 2020; Portuguese funds through FCT—Fundação para a Ciência e a Tecnologia/Ministério da Ciência, Tecnologia e Inovação in the framework of the Decreto-Lei nº 57/2016 de 29 de Agosto (NP work contract) and through the doctoral grant SFRH/BD/136284/2018; and the projects "Institute for Research and Innovation in Health Sciences" (POCI-01-0145-FEDER-007274) and "Center of Mathematics of the University of Porto" (UID/MAT/00144/2019).

LG (ref. 305330/2016-0) was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq and Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro – FAPERJ (CNE-2018).

References

[1] N. Pinto, C. Alves, L. Gusmão, A. Amorim, Theory and statistics of mutation rates: a mathematical framework reformulation for forensic applications, Forensic Sci. Int. Genet. Suppl. Ser. 5 (2015) e131–e132.
[2] N. Pinto, L. Gusmao, A. Amorim, Mutation and mutation rates at Y chromosome specific Short Tandem Repeat Polymorphisms (STRs): a reappraisal, Forensic Sci. Int. Genet. 9 (2014) 20–24.
[3] K. Slooten, F. Ricciardi, Estimation of mutation probabilities for autosomal STR markers, Forensic Sci. Int. Genet. 7 (3) (2013) 337–344.
[4] C. Schlötterer, D. Tautz, Slippage synthesis of simple sequence DNA, Nucleic Acids Res. 20 (2) (1992) 211–215.
[5] M. Strand, T.A. Prolla, R.M. Liskay, T.D. Petes, Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair, Nature 365 (6443) (1993) 274.

[6] J.L. Weber, C. Wong, Mutation of human short tandem repeats, Hum. Mol. Genet. 2 (8) (1993) 1123–1128.
[7] X. Xu, M. Peng, Z. Fang, X. Xu, The direction of microsatellite mutations is dependent upon allele length, Nat. Genet. 24 (4) (2000) 396.
[8] T. Ohta, M. Kimura, A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population, Genet. Res. 22 (2) (1973) 201–204.
[9] A. Di Rienzo, A.C. Peterson, J.C. Garza, A.M. Valdes, M. Slatkin, N.B. Freimer, Mutational processes of simple-sequence repeat loci in human populations, Proc. Natl. Acad. Sci. 91 (8) (1994) 3166–3170.
[10] M. Kimmel, R. Chakraborty, Measures of variation at DNA repeat loci under a general stepwise mutation model, Theor. Popul. Biol. 50 (3) (1996) 345–367.
[11] A.M. Valdes, M. Slatkin, N.B. Freimer, Allele frequencies at microsatellite loci: the stepwise mutation model revisited, Genetics 133 (3) (1993) 737–749.
[12] L.A. Zhivotovsky, M.W. Feldman, Microsatellite variability and genetic distances, Proc. Natl. Acad. Sci. 92 (25) (1995) 11549–11552.
[13] B. Brinkmann, M. Klintschar, F. Neuhuber, J. Hühne, B. Rolf, Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat, Am. J. Hum. Genet. 62 (6) (1998) 1408–1415.
[14] C.R. Primmer, H. Ellegren, N. Saino, A.P. Møller, Directional evolution in germline microsatellite mutations, Nat. Genet. 13 (4) (1996) 391.

2.1.3. Microsatellites' Mutation Modeling Through the Analysis of the Y-Chromosome Transmission: results of a GHEP-ISFG Working Group

Sofia Antão-Sousa[1,2,3,4,§], Leonor Gusmão[4,§], Nidia M. Modesti[5], Sofía Feliziani[5], Marisa Faustino[1,3], Julyana Ribeiro[4], Elizeu Carvalho[4], Vania Pereira[6], Carmen Tomas[6], Marian M. de Pancorbo[7], Miriam Baeta[7], Rashed Alghafri[8], Reem Matar[8], Juan José Builes[9,10], Nair Gouveia[11], German Burgos[12,13], Maria de Lurdes Pontes[14], Adriana Ibarra[15], Claudia Vieira da Silva[16], Rukhsana Parveen[17], Josep Carreras Carbonell[18], António Amorim[1,2,3], Nadia Pinto[1,2,19,*]

[1]Instituto de Investigação e Inovação em Saúde (i3S), Porto, Portugal
[2]Institute of Molecular Pathology and Immunology, University of Porto (IPATIMUP), Porto, Portugal
[3]Faculty of Sciences, University of Porto (FCUP), Porto, Portugal
[4]DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Rio de Janeiro, Brazil
[5]Centro de Genética Forense, Poder Judicial de Córdoba, Argentina
[6]Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark
[7]BIOMICs Research Group, Lascaray Research Center, University of the Basque Country UPV/EHU, Vitoria-Gasteiz, Spain
[8]International Center for Forensic Sciences, Dubai Police G.H.Q., Dubai, United Arab Emirates
[9]GENES SAS Laboratory, Medellín, Colombia
[10]Institute of Biology, University of Antioquia, Medellín, Colombia
[11]Instituto Nacional de Medicina Legal e Ciências Forenses, I.P. / Serviço de Genética e Biologia Forenses, Delegação do Centro, Portugal
[12]Carrera de Enfermería, Facultad de Ciencias de la Salud, Universidad de Las Américas (UDLA), Quito, Ecuador.
[13]Escuela de Medicina, Facultad de Ciencias de la Salud, Universidad de Las Américas (UDLA), Quito, Ecuador.[14]Instituto Nacional de Medicina Legal e Ciências Forenses, I.P. / Serviço de Genética e Biologia Forenses, Delegação do Norte, Portugal
[15]Laboratorio IDENTIGEN, Universidad de Antioquia, Colombia.
[16]Instituto Nacional de Medicina Legal e Ciências Forenses, I.P. / Serviço de Genética e Biologia Forenses, Delegação do Sul, Portugal
[17]Forensic Services Laboratory, Centre for Applied Molecular Biology, University of the Punjab, Lahore, Pakistan
[18]Policia de la Generalitat de Catalunya – Mossos d'Esquadra. Unitat Central del Laboratori Biològic, Barcelona, Spain
[19]Centre of Mathematics of the University of Porto, Porto, Portugal
§ These authors contributed equally to the work
* Corresponding author

In final preparation. For supplementary material please check Appendix B.

Keywords: Y chromosome, mutation, mutation modeling, mutation rate estimation, microsatellites, Y-STRs.

Highlights:

- GHEP-ISFG organized a collaborative study on Y-STR mutations;
- New and published data on 35 Y-STRs were collected and analyzed;
- Marker-specific mutation rates were estimated for 33 Y-STRs;
- A positive correlation between paternal age and mutation was confirmed;
- Per marker mutational (gain/loss, single/multistep) dynamics was quantified.

Abstract

The haploid mode of transmission of the Y chromosome allows the unambiguous identification of which parental allele resulted in which filial one, an unparalleled feature in any other component of the nuclear genome. This characteristic is especially relevant when genotyping consists of amplicon length determination. The Spanish and Portuguese Speaking Working Group of the International Society for Forensic Genetics (GHEP-ISFG) organized a collaborative study on mutations of Y-chromosomal short tandem repeats (Y-STRs). New data from 2,225 father-son duos, from 10 populations worldwide, and data from 44 previously published reports, corresponding to 25,729 duos, were collected and analyzed. Marker-specific mutation rates were estimated for 33 Y-STRs. Although highly variable depending on the analyzed markers, mutations compatible with the gain or loss of a single repeat showed to be 23.2 times more likely than those involving a greater number of repeats. On the other hand, longer alleles showed to be nearly twice more mutable than the shorter ones. Within the subset of longer alleles, the loss of repeats showed to be nearly twice more likely than the gain. Conversely, shorter alleles showed a symmetrical trend, with repeat gains being twofold more frequent than reductions. A positive correlation between the paternal age and the mutation rate was observed, strengthening previous findings. The amount of data used for reaching the different above-mentioned conclusions varied widely, depending on how the data were reported in the different published reports analyzed. This shows a regrettable waste of produced data, purely due to inadequate communication of the results, supporting an urgent need for the elaboration of publication guidelines for mutation studies.

1. Introduction

The Y chromosome provides invaluable data for the study of the biological mechanisms of germinal mutations, with no parallel in any other component of the nuclear genome. This is due to its haploid mode of genetic transmission, which allows the unambiguous identification of which parental allele originated which filial one, whenever father-son duos are analyzed for simple-structure markers [1]. Indeed, when analyzing length polymorphisms for either autosomal or X-chromosomal markers, it is rarely possible to determine with certainty which parental allele originated which filial one, or even if an undetected mutation occurred, as length mutations may not lead to Mendelian incompatibilities. Thus, the standard approach of estimating mutation rates through the number of Mendelian incompatibilities observed necessarily leads to the underestimation of the first for both autosomal and X-chromosomal markers [2]–[6]. It is however noteworthy that this occurs in a smaller extent for X-chromosomal transmission, since in father-daughter and mother-son transmissions the parental and filial allele, respectively, involved in each allele transmission is known [2]. Theoretical approaches have been proposed to mitigate this bias for autosomal markers [3]–[6], and an informatics tool was recently presented [7] aiming more accurate mutation rates estimation.

Short tandem repeats (STRs) are the most used markers in population and forensic genetics, being highly polymorphic due to their relatively high mutation rates. The primary mutation mechanism is thought to be the polymerase template slippage that leads to the insertion or deletion of single or multiple repeats in the allele transmitted from the parent to the child [8], [9].

Concerning STR mutations, and regardless of the mode of genetic transmission considered, it is generally accepted that (*a*) single-step mutations, i.e., the gain or loss of a single repeat, are more frequent than multistep ones [10], [11], (*b*) longer alleles are more prone to mutation than shorter ones [12], and (*c*) longer alleles tend to lose repeats [13]–[16], while shorter ones are more prone to gains than losses [14]–[16]. Moreover, it is also commonly accepted that (*d*) older fathers are more prone to allele mutations (and, for the autosomal and X-linked transmission modes, that (*e*) paternal mutations are more frequent than maternal ones [15], [17]. These premises have been consolidated considering any mode of genetic transmission and mostly considering genotyping methodologies based on fragment length determination, which may lead to biased interpretations, especially when analyzing autosomal or X-chromosomal STRs. For example, a Mendelian incompatibility caused by a multistep mutation may be explained by a single-step one, depending on the genotypic configuration of the parent(s)-child duo or trio analyzed. In this case, since single-step mutations are assumed to be much more frequent than the multistep ones, the first is assumed to have occurred, which

necessarily (and artificially) increases the preponderance of single-step over the multistep mutations. The weight of this bias was previously studied for both autosomal [3]–[7] and X-chromosomal STRs [18], [19].

The study of Y-chromosomal STRs (Y-STRs), specifically those with simple structure, is thus an invaluable approach for STR mutation modeling. Indeed, the unambiguous identification of which parental allele originated which filial one, provides unparalleled insights into the strength of premises (a) to (d) above.

Here, the results of a collaborative study on Y-STR mutations organized by the Spanish and Portuguese Speaking Working Group of the International Society for Forensic Genetics (GHEP-ISFG) is presented, adding new data to the results of a thorough revision of the literature. Besides updating the mutation rates estimates of the most widely used Y-STRs, our approach allowed the quantification of the relative frequencies of single- and multistep mutations, as well as comparing the number of mutations involving shorter and longer (relatively to the modal one) paternal alleles. The trend for paternal alleles to either gain or lose repeats, depending on their length, was also quantified, as well as the correlation between the paternal age and the frequency of mutations.

2. Material and Methods

At the 2019 General Assembly, a GHEP-ISFG collaborative study on Y-STR mutations was approved and opened to the participation of all GHEP-ISFG members. Presentation of the certificates of a proficiency test for the previous three years showing correct results for the analyzed kits was required for each participating laboratory.

2.1. Sample collection and genotyping

New data from 2,225 father/son duos were obtained from 10 worldwide populations: Argentina (414), Brazil (202), Colombia (222), Denmark (96), Ecuador (102), Greenland (104), Pakistan (110), Portugal (509), Spain (250), and United Arab Emirates (216). Each laboratory ensured the anonymization of the samples and total compliance with the legal and ethical requirements for the use of the corresponding data in this research project. Most of the used samples were originally obtained from paternity investigations, where the biological relationship was confirmed using autosomal markers (likelihood ratio

greater than 10^5). Some volunteer duos, with the biological relationship legally recognized, were also analyzed.

Each participating laboratory presented genotypic data regarding a minimum number of father-son duos, analyzed by either PowerPlex® Y23 System, Promega (PPY23), or Yfiler™ Plus PCR Amplification kit, ThermoFisher Scientific (YFPlus).

2.2. Literature review

Worldwide data regarding father-son genotypic configurations for the 27 Y-STRs included in either (or both) PPY23 or YFPlus kits were gathered from 44 published works [20]–[63], as well as for 8 further Y-STRs (not included in the previously mentioned kits): DYS388, DYS435, DYS461, DYS526a/b, DYS547, DYS612, DYS626, and GATA A10.

The total number of father-son duos used in each of the analyzes performed depended on the information available in published works or on our ability to obtain the data produced (Table S1). Namely, marker-specific mutation rates (Results 3.1.) were obtained using data from the present GHEP-ISFG collaborative study and a set of 44 published reports [20]–[63], for which the absolute numbers of analyzed meiosis and observed mutations were presented. The correlation between the age of the fathers at the time of the birth of the son and the mutation frequency (Results 3.2.) was obtained combining the data herein presented with those from two previous GHEP-ISFG collaborative studies [29], [39] and from [59], in which paternal age was reported for both the mutated and unmutated transmissions. Age raw data from [39] were used to match age intervals. Finally, the correlation between the length of the observed alleles and both the occurrence of mutation, and the repeat gain or loss (Results 3.3.) was assessed (and quantified) using data herein presented GHEP-ISFG working commission and from [25], [26], [28], [29], [31], [32], [39], [43], [47]–[50], [59]. The analysed data is herein presented as supplementary material in the format of mutation matrices for a set of 31 Y-STRs (Supplementary Mutation Matrices). Works for which we could not assess the complete set of observed alleles were not considered for this analysis, as the relative length of the mutated alleles compared to the observed ones could not be ascertained.

2.3. Data analysis

The statistical power of the results was assessed considering a level of significance $\alpha=0.05$. Marker-specific mutation rates were estimated for 33 Y-STRs, and the

corresponding confidence intervals were estimated from the binomial standard deviation. The number of mutations compatible with the gain or loss of a specific number of repeats was also assessed.

The correlation between the paternal age and the occurrence of mutation was assessed through Chi-square tests, as well as the correlation between the length of the alleles and both (a.) the occurrence of a mutation, and (b.) the gain or loss of repeats for the 31 Y-STRs with single structure. For analysis (b.), the modal paternal allele, i.e. the allele with the greatest number of observations, was established for each marker. For each marker, paternal alleles were then included into one of the following three categories: *i.* modal allele, *ii.* short, and *iii.* long allele (relative to the modal one). The correlation between the occurrence of mutations compatible with either the gain or loss of repeats, and the length of the alleles was then assessed for each category.

3. Results

3.1. Average mutation rates

The number of father-son duos analyzed for each of the 36 studied markers varied from 24,148 for DYS19, to 161 for DYS435, the median and the average number of pairs of subjects equating 9,976 and 14,154, respectively. Due to the low number of allele transfers analyzed, markers DYS435 and DYS461 (N=161 and N=873, respectively) were disregarded. DYF387S1 was also not considered, since it is a multi-copy marker with a variable number of loci amongst individuals, which does not allow for the correct counting of allele transfers. A total of 467,073 allelic transfers were then analyzed for the remaining 33 Y-STRs, and 1,863 mutations were observed. Average mutation rates varied between 0.0005 (for DYS438 and DYS643) and 0.0170 (for DYS547) (Table 1). Discrimination between allele transmissions and mutations observed in previous works [20]–[63] and those of this study are presented in Table S2. Mutations observed in DYF387S1 are presented in Table S3.

Out of the 1,863 mutations observed, 1,786 were compatible with single-step mutations, and 77 with multistep ones (Table 1), the first being 23.2 times more frequent than the others. It is however noteworthy that this ratio showed to be highly variable between markers, varying between 1.25 (i.e., essentially equifrequent) for DYS438 and 98 for DYS449.

Table 1: Mutation rates estimation for 33 Y-STRs, and corresponding confidence intervals. Number of mutations compatible with either the gain or loss of 1 to 6 repeats, and corresponding ratios between the number of single and multistep mutations. Data include both those generated in the present study and those obtained from literature [20]–[63]. NC: Not compatible with changes involving an integer number of repeats.

Markers	Observations (N)		Mut rate	CI (95%)	Number of repeats compatible to be involved in the mutation							Single/ Multistep
	Muts	**Allele transfers**			**1**	**2**	**3**	**4**	**6**	**>1***	**NC**	
DYS19	57	26,372	0.0022	0.00164-0.00280	56	0	0	0	0	0	1	-
DYS385a/b	112	46,659	0.0024	0.00198-0.00289	104	7	1	0	0	0	0	13.00
DYS388	2	3,612	0.0006	0.00007-0.00200	2	0	0	0	0	0	0	-
DYS389I	69	26,154	0.0026	0.00205-0.00334	69	0	0	0	0	0	0	-
DYS389II-I	91	26,113	0.0035	0.00281-0.00428	88	2	1	0	0	0	0	29.33
DYS390	63	25,103	0.0025	0.00193-0.00321	62	1	0	0	0	0	0	62.00
DYS391	66	25,789	0.0026	0.00198-0.00325	64	2	0	0	0	0	0	32.00
DYS392	14	24,266	0.0006	0.00032-0.00097	13	0	0	1	0	0	0	13.00
DYS393	31	24,332	0.0013	0.00087-0.00181	31	0	0	0	0	0	0	-
DYS437	22	21,018	0.0010	0.00066-0.00158	22	0	0	0	0	0	0	-
DYS438	10	21,033	0.0005	0.00023-0.00087	5	3	0	1	0	0	1	1.25
DYS439	110	21,111	0.0052	0.00428-0.00628	110	0	0	0	0	0	0	-
DYS448	20	16,577	0.0012	0.00074-0.00186	19	1	0	0	0	0	0	19.00
DYS449	99	8,480	0.0117	0.00950-0.01420	98	0	1	0	0	0	0	98.00
DYS456	81	17,892	0.0045	0.00360-0.00562	80	0	1	0	0	0	0	80.00
DYS458	126	17,920	0.0070	0.00586-0.00837	122	3	1	0	0	0	0	30.50
DYS460	23	5,605	0.0041	0.00260-0.00615	23	0	0	0	0	0	0	-
DYS481	25	5,504	0.0045	0.00294-0.00670	22	3	0	0	0	0	0	7.33

Markers	Observations (N)		Mut rate	CI (95%)	Number of repeats compatible to be involved in the mutation							Single/ Multistep
	Muts	**Allele transfers**			**1**	**2**	**3**	**4**	**6**	**>1***	**NC**	
DYS518	116	7,795	0.0149	0.01231-0.01782	105	5	3	2	0	0	1	10.50
DYS526a	15	4,425	0.0034	0.00190-0.00559	11	4	0	0	0	0	0	2.75
DYS526b-a	50	4,401	0.0114	0.00843-0.01498	49	1	0	0	0	0	0	49
DYS533	9	6,975	0.0013	0.00059-0.00245	8	1	0	0	0	0	0	8.00
DYS547	69	4,053	0.0170	0.01327-0.02150	67	0	0	1	0	1	0	33.50
DYS549	10	2,617	0.0038	0.00183-0.00702	10	0	0	0	0	0	0	-
DYS570	87	9,976	0.0087	0.00699-0.01075	80	6	1	0	0	0	0	11.43
DYS576	140	9,876	0.0142	0.01194-0.01671	135	3	2	0	0	0	0	27.00
DYS612	66	4,056	0.0163	0.01261-0.02066	60	6	0	0	0	0	0	10.00
DYS626	41	4,441	0.0092	0.00663-0.01250	39	1	0	0	1	0	0	19.50
DYS627	133	8,028	0.0166	0.01389-0.01960	127	5	1	0	0	0	0	21.17
DYS635	59	18,813	0.0031	0.00239-0.00404	59	0	0	0	0	0	0	-
DYS643	1	1,978	0.0005	0.00001-0.00281	1	0	0	0	0	0	0	-
GATA A10	4	1,026	0.0039	0.00106-0.00995	4	0	0	0	0	0	0	-
GATA H4	42	17,611	0.0024	0.00172-0.00322	41	1	0	0	0	0	0	41.00
Total	1,863	467,073	-	-	1,786	55	12	5	1	1	3	23.2

* Multistep mutation reported without specifying the number of steps [60].

3.2. Correlation between the age of the fathers and the occurrence of mutation

Information on the age of the father at the time of the birth of the son was gathered for 84,715 allele transmissions (Tables 2 and S4). The findings support that the mutation rate and the age of the father are positively correlated. The 10-years age class comprising more subjects was the one with fathers with ages between 21 and 30 years old, gathering 46.5% of the analyzed individuals. The age class 51-60 was the one showing the highest mutation rate (Table 2). Considering dichotomous classes, statistically significant differences were found between individuals aged <30 and >31 (p-value = 0.00302), and <40 and >41 (p-value = 0.00018), and a nearly significant difference was reached between individuals aged <50 and >51 (p-value = 0.05830) (Table S5). In all the cases, the older individuals were associated with higher mutation rates.

Table 2: Number of father-son allele transmissions analyzed considering paternal age intervals (at the time of son's birth) and the corresponding mutation rate. Included data are from the presented study and previous reports [29], [39], [59]. See Tables S4 and S5 for more details.

Paternal age classes		Number of mutations	Number of allelic transmissions	Mutation rate	Confidence interval (95%)
10-years	<21	31	7,995	0.00388	0.00264-0.0055
	21-30	112	39,408	0.00284	0.00234-0.00342
	31-40	96	25,995	0.00369	0.00299-0.00451
	41-50	45	8,076	0.00557	0.00407-0.00745
	51-60	14	2,407	0.00582	0.00318-0.00974
	>61	4	834	0.00480	0.00131-0.01223
Dichotomous	<31	143	47,403	0.00302	0.00254-0.00355
	>30	159	37,312	0.00426	0.00363-0.00498
	<41	239	73,398	0.00326	0.00286-0.00370
	>40	63	11,317	0.00557	0.00428-0.00712
	<51	284	81,474	0.00349	0.00309-0.00392
	>50	18	3,241	0.00555	0.00329-0.00878

3.3. Correlation between the length of the analyzed alleles and mutation occurrence

The number of allele transfers, mutations, and mutations compatible with either the gain or loss of one or more repeats were analyzed for each marker, considering the following categories for the observed paternal alleles: modal allele, either shorter or longer than the modal allele (Table 3). The presented results were reached for 31 Y-STRs through the mutation matrices presented in Supplementary Material that were elaborated considering the data obtained in this study and from [25], [26], [28], [29], [31], [32], [39], [43], [47]–[50], [59]. Two markers from the set of those analysed in section 3.1 were excluded from the analysis: DYS385a/b because the identification of mutations would result ambiguous, and DYS643 because no mutations were reported for this marker in the data used for this analysis.

Considering the subset of 62,378 paternal alleles shorter than the modal one, 241 suffered mutation, which resulted in an overall mutation rate of 3.7E-03. This figure increases to 9.8E-03 if considering the 621 mutations observed among the 63,115 transfers involving paternal alleles greater than the modal one. Indeed, longer alleles showed to be 2.55 times more prone to mutation than shorter ones, this difference showing strong statistical significance (p=1.36E-37).

On the other hand, when considering the gain or loss of repeats, most of the mutations involving shorter paternal alleles (241) involved the gain of repeats (161), this difference between gains and losses showing statistical significance (p=1. 82E-04). Conversely, 394 out of the 621 mutations involving longer paternal alleles corresponded to the loss of repeats, this difference between gains and losses also showing statistical significance (p=8.4E-06). The trend for shorter alleles to gain repeats and longer ones to lose repeats showed to be highly significant (p=8.92E-15). Finally, it should be remarked that mutations involving the modal allele did not show statistically supported evidence to either gain or lose repeats (p= 0.11).

Table 3: Number of allele transfers (Transf) and mutations observed compatible with changes involving an interger number of repeats (Mut), as well as number of Mut compatible with either gains or losses, considering the following categories for the paternal allele: modal allele, shorter or longer than de modal allele. Data include both those generated in the presented GHEP-ISFG collaborative study and those obtained from [25], [26], [28], [29], [31], [32], [39], [43], [47]–[50], [59].

Markers	Transf	Mut	Modal allele	Length of Paternal allele analysed (relative to the modal)											
				Shorter than modal				Modal				Longer than modal			
				Transf	Mut	Loss	Gain	Transf	Mut	Loss	Gain	Transf	Mut	Loss	Gain
DYS19	11,688	26	14	1,463	2	0	2	5,333	10	1	9	4,892	14	7	7
DYS388	101	1	12	19	1	0	1	73	0	0	0	9	0	0	0
DYS389I	10,616	37	13	2,426	6	1	5	6,011	14	4	10	2,179	17	15	2
DYS389II-I	10,593	48	16	886	4	0	4	5,456	10	0	10	4,251	34	23	21
DYS390	11,825	29	24	4,643	5	1	4	5,174	15	8	7	2,008	9	9	0
DYS391	11,827	35	10	632	1	0	1	6,775	10	1	9	4,420	24	14	10
DYS392	11,810	7	13	5,537	1	1	0	4,794	3	0	3	1,479	3	1	2
DYS393	10,577	17	13	2,131	2	0	2	6,848	8	4	4	1,598	7	5	2
DYS437	10,057	15	14	37	0	0	0	4,720	3	0	3	5,300	12	8	4
DYS438	10,119	6	12	6,340	4	3	1	3,435	2	2	0	344	0	0	0
DYS439	10,078	63	12	4,330	13	3	10	4,258	18	6	12	1,490	32	23	9
DYS448	8,756	13	19	1,083	1	0	1	3,436	6	4	2	4,237	6	6	0
DYS449	5,538	69	30	1,769	12	4	8	952	11	8	3	2,817	46	25	21
DYS456	8,757	53	15	1,213	3	1	2	4,005	14	2	12	3,539	36	25	11
DYS458	8,765	79	17	3,884	23	8	15	2,620	24	12	12	2,261	32	18	14
DYS460	3,411	14	11	1,797	5	1	4	1,466	5	4	1	148	4	4	0

Markers	Transf	Mut	Modal allele	Length of Paternal allele analysed (relative to the modal)											
				Shorter than modal				Modal				Longer than modal			
				Transf	Mut	Loss	Gain	Transf	Mut	Loss	Gain	Transf	Mut	Loss	Gain
DYS481	3,705	21	22	275	0	0	0	1,221	5	1	4	2,209	16	6	10
DYS518	5,493	93	38	1,425	14	4	10	1,166	15	7	8	2,902	64	32	32
DYS526a	3,119	14	14	890	6	3	3	1,000	4	2	2	1,229	4	2	2
DYS526b-a	3,095	38	22	842	6	2	4	773	8	1	7	1,480	24	12	12
DYS533	3,705	5	12	1,718	1	0	1	1,714	2	1	1	273	2	2	0
DYS547	3,259	57	48	1,223	21	8	13	778	9	4	5	1,258	27	22	5
DYS549	1,317	7	12	240	0	0	0	491	3	2	1	586	4	4	0
DYS570	7,015	64	17	899	7	0	7	2,060	18	6	12	4,056	39	25	14
DYS576	6,914	108	18	3,111	20	3	17	2,191	35	20	15	1,612	53	37	16
DYS612	3,288	57	36	884	21	5	16	933	6	3	3	1,471	30	15	15
DYS626	3,237	35	30	1,241	6	3	3	610	2	1	1	1,386	27	20	7
DYS627	5,663	89	21	2,811	28	10	18	1,152	15	9	6	1,700	46	29	17
DYS635	9,635	35	23	4,424	20	16	4	4,141	9	6	3	1,070	6	3	3
GATA A10	874	4	15	316	2	1	1	433	2	2	0	125	0	0	0
GATA H4	9,848	22	12	3,889	6	2	4	5,173	13	9	4	786	3	2	1
Overall average	**214,685**	**1,161**	**-**	**62,378**	**241**	**80**	**161**	**89,192**	**299**	**130**	**169**	**63,115**	**621**	**394**	**237**

4. Discussion and Conclusions

The analysis of the transmission of Y-chromosomal markers provides invaluable insights into mutation mechanisms, especially when genotyping is based on fragment length determination, as it is still the common practice in forensic genetics routine. Indeed, the unambiguous identification in simple-structure STRs of which parental allele originated which filial one, prevents the biases inherent to other modes of genetic transmission, such as the occurrence of hidden mutations or the overestimation of single-step mutations compared to multi-step ones [3], [7], [18]. The insights provided by these haploid markers may be analyzed under the scope of autosomal and X-chromosomal modes of genetic transmission, and conclusions eventually transferred and included in specifically devoted software. It must be said, however that the absence of recombination at Y specific region may limit the transferability of the model.

The average mutation rates of the 33 different Y-STRs vary significantly, from 0.0005 for DYS438 and DYS643 to 0.0170 for DYS547 (i.e., 34 times), which strengthens the recommendation on the use of marker-specific estimates. For markers with complex structure the presented mutation rate may be underestimated if 'compensating' mutations occur within the markers (that is, a gain in one region and an equal length loss in another).

Our work also supports that single-step mutations are more frequent than multistep ones, although the ratio between the two varies from marker to marker – from slightly above 1 (DYS438) to almost one hundred (DYS449).

Longer alleles (relative to the modal one) showed ~twice more mutations than the shorter ones. Within the subset of longer alleles, the trend to lose repeats showed to be greater (1.7 times) than for gains, the opposite trend being observed within the subset of shorter alleles (2.0 times more mutations involving gains than losses). When considering the mutations involving the modal paternal allele, no statistically significant differences were found between the numbers of the two types of mutations.

In agreement with previous reports [15], [16], [64], [65], the data analyzed in this paper support a positive correlation between the age of the father and the occurrence of mutation.

In summary, our results provide quantified evidence supporting the generally accepted premises that (***i.***) single-step mutations are more common than multistep ones, but showed that the magnitude of the difference is highly variable across the analyzed

markers, (***ii.***) long alleles are more prone to mutation than short ones (~ twice), and (***iii.)*** long alleles are more prone to lose repeats (~ twice) while short alleles show the opposite tendency (also ~ twice).

In any case, it is noteworthy that the statistical strength of the conclusions described above is greatly dependent on the level of detail under which the data are published. Indeed, 44 published reports were usable to estimate the average mutation rates presented in Table 1 (467,073 father-son allele transmissions), as all of them reported for each marker both the number of allele transmissions and the number of mutations observed [20]–[63]. The age of the father at the time of the birth of the child was obtained only for 84,715 allele transmissions out of the 467,073 produced and analyzed (18.1%). On the other hand, the complete set of analyzed allele transmissions, including non-mutated alleles, which allow us to present Table 3, was only possible to gather for 214,685 out of the 467,073 allele transmissions produced and analyzed regarding single copy markers and compatible with either the absence of mutation or changes involving an integer number of repeats. This represents a huge amount of wasted, or at least under-exploited, data (46.0%), which would be useful to advance the state of the art of microsatellites' mutation modeling. Specifically, a proper communication of the produced data would allow a more robust estimation of bi-allele mutation rates, essential for the weighing of the evidence in kinship problems [1]. Indeed, there is an obvious need for publication guidelines regarding mutation data, as previously remarked [64], a standardization that would allow produced data to be (re)used by other authors.

Proper estimation of mutation parameters is crucial for a wide range of forensic genetic problems as well as in evolutionary and phylogenetic studies. Since mutation is a rare event, proper modeling necessarily depends on the analysis of a number of individuals prohibitively large for being collected and analyzed by a single laboratory. The organization of collaborative studies gathering efforts and synergies from several laboratories, such as the one here presented from the GHEP-ISFG, seems a good strategy to overcome the difficulty of recruiting such a large amount of genetic data. Properly produced, analyzed and peer-reviewed data must be suitably communicated in a way that may be reused, verified, and re-analyzed later for further investigations. The only way to report data allowing the study of the co-occurrence of mutations in the same meiosis (an overlooked possibility deserving investigation) would require data release in haplotypic format. We recognize however that the public release of individual haplotypes, even anonymized, may raise ethical concerns. Anyhow we urge the forensic community and their representative bodies to undertake the elaboration of recommendations concerning the publication of results on this topic. In accordance we dare to suggest as

a minimum standard the mutation matrix format used in this work and presented as supplementary material which, although not allowing the investigation of co-occurrence of mutations, provides nonetheless the information required for the estimation of per marker, allele specific, mutation rates.

Acknowledgments

We deeply thank the support of Promega and ThermoFisher Scientific as reduced prices were provided for the participants in the purchase of the kits PowerPlex® Y23 System, Promega (PPY23), and Yfiler™ Plus PCR Amplification kit, ThermoFisher Scientific (YFPlus), respectively. This work was partially financed by FEDER- Fundo Europeu de Desenvolvimento Regional funds through the COMPETE - Operational Program for Competitiveness and Internationalization POCI, Portugal, and by Portuguese funds through FCT- Fundação para a Ciência e a Tecnologia, Ministério da Ciência, Tecnologia e Inovação in the framework of the projects "Institute for Research and Innovation in Health Sciences" POCI-01-0145- FEDER-007274). NP is supported by FCT, under the program contract provided in Decree-Law no.57/2016 of August 29. SAS is funded by a FCT doctoral grant SFRH/BD/136284/2018. LG is supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq ref. 306342/2019-7), and Fundação de Amparo Pesquisa do Estado do Rio de Janeiro - FAPERJ (CNE-2022).

References

[1] N. Pinto, L. Gusmão, and A. Amorim, "Mutation and mutation rates at Y chromosome specific Short Tandem Repeat Polymorphisms (STRs): a reappraisal," *Forensic Sci Int Genet*, vol. 9, no. 1, pp. 20–24, Mar. 2014, doi: 10.1016/J.FSIGEN.2013.10.008.

[2] S. Antão-Sousa, E. Conde-Sousa, L. Gusmão, A. Amorim, and N. Pinto, "Underestimation and misclassification of mutations at X chromosome STRs depend on population's allelic profile," *Forensic Sci Int Genet Suppl Ser*, vol. 7, no. 1, pp. 718–720, Dec. 2019, doi: 10.1016/J.FSIGSS.2019.10.150.

[3] K. Slooten and F. Ricciardi, "Estimation of mutation probabilities for autosomal STR markers," *Forensic Sci Int Genet*, vol. 7, no. 3, pp. 337–344, May 2013, doi: 10.1016/J.FSIGEN.2013.01.006.

[4] R. Chakraborty, D. N. Stivers, and Y. Zhong, "Estimation of mutation rates from parentage exclusion data: applications to STR and VNTR loci," *Mutat Res*, vol. 354, no. 1, pp. 41–48, Jul. 1996, doi: 10.1016/0027-5107(96)00014-0.

[5] P. Vicard and A. P. Dawid, "A statistical treatment of biases affecting the estimation of mutation rates," *Mutat Res*, vol. 547, no. 1–2, pp. 19–33, Mar. 2004, doi: 10.1016/J.MRFMMM.2003.11.005.

[6] P. Vicard, A. P. Dawid, J. Mortera, and S. L. Lauritzen, "Estimating mutation rates from paternity casework," *Forensic Sci Int Genet*, vol. 2, no. 1, pp. 9–18, Jan. 2008, doi: 10.1016/J.FSIGEN.2007.07.002.

[7] S. Antão-Sousa, E. Conde-Sousa, L. Gusmão, A. Amorim, and N. Pinto, "Estimations of Mutation Rates Depend on Population Allele Frequency Distribution: The Case of Autosomal Microsatellites," *Genes (Basel)*, vol. 13, no. 7, Jul. 2022, doi: 10.3390/genes13071248.

[8] M. Strand, T. A. Prolla, R. M. Liskay, and T. D. Petes, "Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair," *Nature 1993 365:6443*, vol. 365, no. 6443, pp. 274–276, 1993, doi: 10.1038/365274a0.

[9] C. Schlötterer and D. Tautz, "Slippage synthesis of simple sequence DNA.," *Nucleic Acids Res*, vol. 20, no. 2, p. 211, Jan. 1992, doi: 10.1093/NAR/20.2.211.

[10] X. Xu, M. Peng, Z. Fang, and X. Xu, "The direction of microsatellite mutations is dependent upon allele length," *Nat Genet*, vol. 24, no. 4, pp. 396–399, Apr. 2000, doi: 10.1038/74238.

[11] J. L. Weber and C. Wong, "Mutation of human short tandem repeats," *Hum Mol Genet*, vol. 2, no. 8, pp. 1123–1128, Aug. 1993, doi: 10.1093/HMG/2.8.1123.

[12] M. Wierdl, M. Dominska, and T. D. Petes, "Microsatellite instability in yeast: dependence on the length of the microsatellite," *Genetics*, vol. 146, no. 3, pp. 769–779, 1997, doi: 10.1093/GENETICS/146.3.769.

[13] H. Ellegren, "Heterogeneous mutation processes in human microsatellite DNA sequences," 2000. [Online]. Available: http://www.ebc.uu.se/evbiol/index.shtml

[14] S. Antão-Sousa, A. Amorim, L. Gusmão, and N. Pinto, "Mutation in Y STRs: Repeat motif gains vs. losses," *Forensic Sci Int Genet Suppl Ser*, vol. 7, no. 1, pp. 240–242, Dec. 2019, doi: 10.1016/j.fsigss.2019.09.092.

[15] J. X. Sun *et al.*, "A direct characterization of human mutation based on microsatellites," *Nat Genet*, vol. 44, no. 10, pp. 1161–1165, Oct. 2012, doi: 10.1038/NG.2398.

[16] K. N. Ballantyne *et al.*, "Mutability of Y-Chromosomal Microsatellites: Rates, Characteristics, Molecular Bases, and Forensic Implications," *Am J Hum Genet*, vol. 87, no. 3, p. 341, Sep. 2010, doi: 10.1016/J.AJHG.2010.08.006.

[17] N. Pinto *et al.*, "Paternal and maternal mutations in X-STRs: A GHEP-ISFG collaborative study," *Forensic Sci Int Genet*, vol. 46, May 2020, doi: 10.1016/J.FSIGEN.2020.102258.

[18] S. Antão-Sousa, E. Conde-Sousa, L. Gusmão, A. Amorim, and N. Pinto, "How frequently are Autosomal and X-STRs multistep mutations perceived as single-step?," *Forensic Sci Int Genet Suppl Ser*, Dec. 2022, doi: 10.1016/j.fsigss.2022.10.022.

[19] S. Antão-Sousa, E. Conde-Sousa, L. Gusmão, A. Amorim, and N. Pinto, "How often have X- and autosomal-STRs mutations equivocal parental origin been assigned?," *Forensic Sci Int Genet Suppl Ser*, 2022, doi: 10.1016/j.fsigss.2022.09.035.

[20] R. Lessig and J. Edelmann, "Y chromosome polymorphisms and haplotypes in west Saxony (Germany)," *Int J Legal Med*, vol. 111, no. 4, pp. 215–218, Jun. 1998, doi: 10.1007/S004140050155.

[21] N. O. Blanchi *et al.*, "Characterization of Ancestral and Derived Y-Chromosome Haplotypes of New World Native Populations," *The American Journal of Human Genetics*, vol. 63, no. 6, pp. 1862–1871, Dec. 1998, doi: 10.1086/302141.

[22] M. Kayser *et al.*, "Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs.," *Am J Hum Genet*, vol. 66, no. 5, p. 1580, 2000, doi: 10.1086/302905.

[23] B. M. Dupuy *et al.*, "Y-chromosome variation in a Norwegian population sample," *Forensic Sci Int*, vol. 117, no. 3, pp. 163–173, Apr. 2001, doi: 10.1016/S0379-0738(00)00397-2.

[24] B. M. Dupuy, M. Stenersen, T. Egeland, and B. Olaisen, "Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci," *Hum Mutat*, vol. 23, no. 2, pp. 117–124, 2004, doi: 10.1002/HUMU.10294.

[25] L. C. Tsai *et al.*, "Haplotype frequencies of nine Y-chromosome STR loci in the Taiwanese Han population," *Int J Legal Med*, vol. 116, no. 3, pp. 179–183, 2002, doi: 10.1007/S004140100236.

[26] R. Kurihara *et al.*, "Mutations in 14 Y-STR loci among Japanese father-son haplotypes," *Int J Legal Med*, vol. 118, no. 3, pp. 125–131, Jun. 2004, doi: 10.1007/S00414-003-0422-3.

[27] B. Budowle *et al.*, "Twelve short tandem repeat loci Y chromosome haplotypes: genetic analysis on populations residing in North America," *Forensic Sci Int*, vol. 150, no. 1, pp. 1–15, May 2005, doi: 10.1016/J.FORSCIINT.2005.01.010.

[28] B. Berger, A. Lindinger, H. Niederstätter, P. Grubwieser, and W. Parson, "Y-STR typing of an Austrian population sample using a 17-loci multiplex PCR assay," *Int J Legal Med*, vol. 119, no. 4, pp. 241–246, 2005, doi: 10.1007/S00414-005-0546-8.

[29] L. Gusmão *et al.*, "Mutation rates at Y chromosome specific microsatellites," *Hum Mutat*, vol. 26, no. 6, pp. 520–528, Dec. 2005, doi: 10.1002/HUMU.20254.

[30] D. J. Ballard *et al.*, "A study of mutation rates and the characterisation of intermediate, null and duplicated alleles for 13 Y chromosome STRs," *Forensic Sci Int*, vol. 155, no. 1, pp. 65–70, Dec. 2005, doi: 10.1016/J.FORSCIINT.2004.12.012.

[31] S. Turrina, R. Atzei, and D. de Leo, "Y-chromosomal STR haplotypes in a Northeast Italian population sample using 17plex loci PCR assay," *Int J Legal Med*, vol. 120, no. 1, pp. 56–59, Jan. 2006, doi: 10.1007/S00414-005-0054-X.

[32] H. Y. Lee *et al.*, "Haplotypes and mutation analysis of 22 Y-chromosomal STRs in Korean father-son pairs," *Int J Legal Med*, vol. 121, no. 2, pp. 128–135, Mar. 2007, doi: 10.1007/S00414-006-0130-X.

[33] C. Hohoff, K. Dewa, U. Sibbing, K. Hoppe, P. Forster, and B. Brinkmann, "Y-chromosomal microsatellite mutation rates in a population sample from northwestern Germany," *Int J Legal Med*, vol. 121, no. 5, pp. 359–363, Sep. 2007, doi: 10.1007/S00414-006-0123-9.

[34] R. Kumagai, A. Kumagai, K. Saigusa, and Y. Aoki, "Haplotype analysis of 17 Y-STR loci in a Japanese population," *Forensic Sci Int*, vol. 172, no. 1, pp. 72–78, Oct. 2007, doi: 10.1016/J.FORSCIINT.2006.07.022.

[35] M. L. Pontes, L. Cainé, D. Abrantes, G. Lima, and M. F. Pinheiro, "Allele frequencies and population data for 17 Y-STR loci (AmpFlSTR Y-filer) in a Northern Portuguese population sample," *Forensic Sci Int*, vol. 170, no. 1, pp. 62–67, Jul. 2007, doi: 10.1016/J.FORSCIINT.2006.04.008.

[36] M. sen Shi, J. P. Tang, R. F. Bai, X. J. Yu, J. Y. Lv, and B. Hu, "Haplotypes of 20 Y-chromosomal STRs in a population sample from southeast China (Chaoshan area)," *Int J Legal Med*, vol. 121, no. 6, pp. 455–462, Nov. 2007, doi: 10.1007/S00414-007-0186-2.

[37] P. M. Domingues, L. Gusmão, D. A. da Silva, A. Amorim, R. W. Pereira, and E. F. de Carvalho, "Sub-Saharan Africa descendents in Rio de Janeiro (Brazil): population and mutational data for 12 Y-STR loci," *Int J Legal Med*, vol. 121, no. 3, pp. 238–241, May 2007, doi: 10.1007/S00414-007-0154-X.

[38] A. E. Decker, M. C. Kline, J. W. Redman, T. M. Reid, and J. M. Butler, "Analysis of mutations in father-son pairs with 17 Y-STR loci," *Forensic Sci Int Genet*, vol. 2, no. 3, Jun. 2008, doi: 10.1016/J.FSIGEN.2007.08.016.

[39] P. Sánchez-Diz *et al.*, "Population and segregation data on 17 Y-STRs: results of a GEP-ISFG collaborative study," *Int J Legal Med*, vol. 122, no. 6, pp. 529–533, Nov. 2008, doi: 10.1007/S00414-008-0265-Z.

[40] J. R. Padilla-Gutiérrez *et al.*, "Population data and mutation rate of nine Y-STRs in a mestizo Mexican population from Guadalajara, Jalisco, México," *Leg Med (Tokyo)*, vol. 10, no. 6, pp. 319–320, Nov. 2008, doi: 10.1016/J.LEGALMED.2008.03.004.

[41] J. A. Soares-Vieira, A. E. C. Billerbeck, E. S. M. Iwamura, B. B. Mendonca, L. Gusmão, and P. A. Otto, "Population and mutation analysis of Y-STR loci in a sample from the city of São Paulo (Brazil)," *Genet Mol Biol*, vol. 31, no. 3, pp. 651–656, 2008, doi: 10.1590/S1415-47572008000400009.

[42] M. Goedbloed *et al.*, "Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFlSTR Yfiler PCR amplification kit," *Int J Legal Med*, vol. 123, no. 6, pp. 471–482, Nov. 2009, doi: 10.1007/S00414-009-0342-Y.

[43] J. Ge, B. Budowle, X. G. Aranda, J. v. Planz, A. J. Eisenberg, and R. Chakraborty, "Mutation rates at Y chromosome short tandem repeats in Texas populations," *Forensic Sci Int Genet*, vol. 3, no. 3, pp. 179–184, Jun. 2009, doi: 10.1016/J.FSIGEN.2009.01.007.

[44] S. H. Kim *et al.*, "Population genetics and mutational events at 6 Y-STRs in Korean population," *Forensic Sci Int Genet*, vol. 3, no. 2, Mar. 2009, doi: 10.1016/J.FSIGEN.2008.06.004.

[45] C. Vieira-Silva, P. Dario, T. Ribeiro, I. Lucas, H. Geada, and R. Espinheira, "Y-STR mutational rates determination in South Portugal Caucasian population," *Forensic Sci Int Genet Suppl Ser*, vol. 2, no. 1, pp. 60–61, Dec. 2009, doi: 10.1016/J.FSIGSS.2009.08.159.

[46] A. Laouina *et al.*, "Mutation rate at 17 Y-STR loci in 'Father/Son' pairs from moroccan population," *Leg Med*, vol. 15, no. 5, pp. 269–271, Sep. 2013, doi: 10.1016/J.LEGALMED.2013.03.003.

[47] K. N. Ballantyne *et al.*, "Toward male individualization with rapidly mutating y-chromosomal short tandem repeats," *Hum Mutat*, vol. 35, no. 8, pp. 1021–1032, 2014, doi: 10.1002/HUMU.22599.

[48] C. Robino *et al.*, "Development of an Italian RM Y-STR haplotype database: Results of the 2013 GEFI collaborative exercise," *Forensic Sci Int Genet*, vol. 15, pp. 56–63, 2015, doi: 10.1016/J.FSIGEN.2014.10.008.

[49] Y. Wang *et al.*, "Genetic polymorphisms and mutation rates of 27 Y-chromosomal STRs in a Han population from Guangdong Province, Southern China," *Forensic Sci Int Genet*, vol. 21, pp. 5–9, Mar. 2016, doi: 10.1016/J.FSIGEN.2015.09.013.

[50] S. Antão-Sousa *et al.*, "Mutation rates and segregation data on 16 Y-STRs: An update to previous GHEP-ISFG studies," *Forensic Sci Int Genet Suppl Ser*, vol. 6, pp. e601–e602, Dec. 2017, doi: 10.1016/J.FSIGSS.2017.10.008.

[51] A. Adnan, A. Rakha, O. Lao, and M. Kayser, "Mutation analysis at 17 Y-STR loci (Yfiler) in father-son pairs of male pedigrees from Pakistan," *Forensic Sci Int Genet*, vol. 36, pp. e17–e18, Sep. 2018, doi: 10.1016/J.FSIGEN.2018.07.001.

[52] F. C. Bugoye, E. Mulima, and G. Misinzo, "Analysis of Mutation Rate of 17 Y-Chromosome Short Tandem Repeats Loci Using Tanzanian Father-Son Paired Samples," *Genet Res Int*, vol. 2018, 2018, doi: 10.1155/2018/8090469.

[53] E. Mertoglu, G. Filoglu, T. Zorlu, and O. Bulbul, "Estimation of the Y-chromosomal short tandem repeat (Y-STR) mutation rates in Turkey," *Turkish Journal of Biochemistry*, vol. 43, no. 2, pp. 142–150, Apr. 2018, doi: 10.1515/TJB-2017-0240/PDF.

[54] Y. Yang *et al.*, "Haplotypic polymorphisms and mutation rate estimates of 22 Y-chromosome STRs in the Northern Chinese Han father–son pairs," *Scientific Reports 2018 8:1*, vol. 8, no. 1, pp. 1–6, May 2018, doi: 10.1038/s41598-018-25362-3.

[55] W. Wu *et al.*, "Mutation rates at 42 Y chromosomal short tandem repeats in Chinese Han population in Eastern China," *Int J Legal Med*, vol. 132, no. 5, pp. 1317–1319, Sep. 2018, doi: 10.1007/S00414-018-1784-X.

[56] V. Petrovic, M. Kecmanovic, M. Keckarevic Markovic, and D. Keckarevic, "Assessment of mutation rates for PPY23 Y chromosome STR loci in Serbian father-son pairs," *Forensic Sci Int Genet*, vol. 39, pp. e5–e9, Mar. 2019, doi: 10.1016/J.FSIGEN.2018.11.014.

[57] J. Zhang *et al.*, "Mutation rates in father-son pairs of the 27 Y-STR loci in the Dezhou Han population from Shandong province, eastern China," *J Forensic Leg Med*, vol. 67, pp. 61–63, Oct. 2019, doi: 10.1016/J.JFLM.2019.06.015.

[58] M. Ay, A. Serin, H. Sevay, C. Gurkan, and H. Canan, "Genetic characterisation of 13 rapidly mutating Y-STR loci in 100 father and son pairs from South and East Turkey," *https://doi.org/10.1080/03014460.2018.1559353*, vol. 45, no. 6–8, pp. 506–515, Nov. 2019, doi: 10.1080/03014460.2018.1559353.

[59] I. B. Ambrosio *et al.*, "Mutational data and population profiling of 23 Y-STRs in three Brazilian populations," *Forensic Sci Int Genet*, vol. 48, Sep. 2020, doi: 10.1016/J.FSIGEN.2020.102348.

[60] H. Lin, Q. Ye, P. Tang, T. Mo, X. Yu, and J. Tang, "Analyzing genetic polymorphism and mutation of 44 Y-STRs in a Chinese Han population of Southern China," *Leg Med (Tokyo)*, vol. 42, Feb. 2020, doi: 10.1016/J.LEGALMED.2019.101643.

[61] J. Fu, J. Cheng, C. Wei, M. A. Khan, Z. Jin, and J. Fu, "Assessing 23 Y-STR loci mutation rates in Chinese Han father-son pairs from southwestern China," *Mol Biol Rep*, vol. 47, no. 10, pp. 7755–7760, Oct. 2020, doi: 10.1007/S11033-020-05851-3.

[62] S. Bredemeyer, L. Roewer, and S. Willuweit, "Next generation sequencing of Y-STRs in father-son pairs and comparison with traditional capillary electrophoresis," *Forensic Sci Res*, 2021, doi: 10.1080/20961790.2021.1898078/SUPPL_FILE/TFSR_A_1898078_SM7293.XLSX.

[63] L. Yuan *et al.*, "Mutation analysis of 13 RM Y-STR loci in Han population from Beijing of China," *Int J Legal Med*, vol. 133, no. 1, pp. 59–63, Jan. 2019, doi: 10.1007/S00414-018-1949-7.

[64] S. Claerhout *et al.*, "Determining Y-STR mutation rates in deep-routing genealogies: Identification of haplogroup differences," *Forensic Sci Int Genet*, vol. 34, pp. 1–10, May 2018, doi: 10.1016/J.FSIGEN.2018.01.005.

[65] A. Ralf *et al.*, "Identification and characterization of novel rapidly mutating Y-chromosomal short tandem repeat markers," *Hum Mutat*, vol. 41, no. 9, pp. 1680–1696, Sep. 2020, doi: 10.1002/humu.24068.

Appendix B

Antão-Sousa, S. & Gusmão, L., *et al.* (2023). Microsatellites' mutation modeling through the analysis of the Y-chromosome transmission: results of a GHEP-ISFG Collaborative Study

Table S1: Analytical description of the publications from Y-STRs mutation data were gathered.

Reference	Markers																																Number of father-son duos used to:			
	DYS19 [1,2]	DYS385a/b* [1,2]	DYS388	DYS389I/II [1,2]	DYS390 [1,2]	DYS391 [1,2]	DYS392 [1,2]	DYS393 [1,2]	DYS435	DYS437 [1,2]	DYS438 [1,2]	DYS439 [1,2]	DYS448 [1,2]	DYS449 [2]	DYS456 [1,2]	DYS458 [1,2]	DYS460 [2]	DYS461	DYS481 [1,2]	DYS518 [2]	DYS526a/b*	DYS533 [1,2]	DYS547	DYS549 [1]	DYS570 [1,2]	DYS576 [1,2]	DYS612	DYS626	DYS627 [2]	DYS635 (C4) [1,2]	DYS643 [1]	GATA A10	GATA H4 [1,2]	DYF387S1 **[2]	Average mutation rates analyses (Results 3.1.)	Age analyses (Results 3.2.)	Alelle length analyses (Results 3.3.)	
																																				Repeats gain or loss	Single or multistep mutations
Lessig & Edelmann [20]	X	X		X	X																													41	0	0	0
Blanchi *et al.* [21]	X			X	X	X	X	X																										249	0	0	0
Kayser *et al.* [22]	X	X		X	X	X	X	X																										415 to 966	0	0	0
Dupuy *et al.* [23]	X	X	X	X	X	X	X	X																										150	0	0	0
Tsai *et al.* [25]	X	X	X	X	X	X	X	X																										109	0	101	101
Kurihara *et al.* [26]	X	X		X	X	X	X	X	X	X	X	X					X															X		161	0	147	147
Dupuy *et al.* [24]	X	X	X	X	X	X	X	X																										1766	0	0	0
Budowle *et al.* [27]	X	X		X	X	X	X	X		X	X	X																						692	0	0	0
Berger *et al.* [28]	X	X		X	X	X	X	X		X	X	X	X		X	X														X			X		70	0	70	70
Gusmão *et al.* [29]	X	X		X	X	X	X	X		X	X	X					X	X												X		X	X	873 to 2816	1355	873 to 2816	873 to 2816
Ballard *et al.* [30]	X	X		X	X	X	X	X		X	X	X					X															X		72 to 249	0	0	0
Turrina *et al.* [31]	X	X		X	X	X	X	X		X	X	X	X		X	X														X			X		104	0	104	104
Lee *et al.* [32]	X	X		X	X	X	X	X		X	X	X	X		X	X														X			X		369	0	355	355
Hohoff *et al.* [33]	X	X		X	X	X	X	X		X	X	X																							1027	0	0	0
Kumagai *et al.* [34]	X	X		X	X	X	X	X			X	X			X	X																			10	0	0	0

Reference	Markers																															Number of father-son duos used to:			
	DYS19 [1, 2]	DYS385a/b* [1, 2]	DYS388	DYS389I/II [1, 2]	DYS390 [1, 2]	DYS391 [1, 2]	DYS392 [1, 2]	DYS393 [1, 2]	DYS435	DYS437 [1, 2]	DYS438 [1, 2]	DYS439 [1, 2]	DYS448 [1, 2]	DYS449 [2]	DYS456 [1, 2]	DYS458 [1, 2]	DYS460 [2]	DYS461 [1, 2]	DYS481 [1, 2]	DYS518 [2]	DYS526a/b*	DYS533 [1, 2]	DYS547	DYS549 [1]	DYS570 [1, 2]	DYS576 [1, 2]	DYS612	DYS626	DYS627 [2]	DYS635 (C4) [1, 2]	DYS643 [1]	GATA A10	GATA H4 [1, 2]	DYF387S1 ** [2]	Average mutation rates analyses (Results 3.1.)	Age analyses (Results 3.2.)	Alelle length analyses (Results 3.3.)	
																																			Repeats gain or loss	Single or multistep mutations
Pontes *et al.* [35]	X	X		X	X	X	X	X		X	X	X	X		X	X													X			X		45	0	0	0
Shi *et al.* [36]											X	X	X		X	X	X												X		X	X		80	0	0	0
Domingues *et al.* [37]	X	X		X	X	X	X	X		X	X	X																						135	0	0	0
Decker *et al.* [38]	X	X		X	X	X	X	X		X	X	X	X		X	X													X			X		389	0	0	0
Sánchez-Diz *et al.* [39]	X	X		X	X	X	X	X		X	X	X	X		X	X													X			X		797	677	797	797
Padilla-Gutiérrez *et al.* [40]	X				X	X	X						X		X	X																		187 to 218	0	0	0
Soares-Vieira *et al.* [41]	X			X	X	X	X	X																										222	0	0	0
Goedbloed *et al.* [42]	X	X		X	X	X	X	X		X	X	X	X		X	X													X			X		1730 to 1762	0	0	0
Ge *et al.* [43]	X	X		X	X	X	X	X		X	X	X	X		X	X													X			X		2813 to 2927	0	99	99
Kim *et al.* [44]																						X		X	X	X				X				152	0	0	0
Vieira-Silva *et al.* [45]	X	X		X	X	X	X	X		X	X	X	X		X	X													X			X		95	0	0	0
Laouina *et al.* [46]	X	X		X	X	X	X	X		X	X	X	X		X	X													X			X		252	0	0	0
Ballantyne *et al.* [47]	X	X		X	X	X	X	X		X	X	X	X	X	X	X				X	X		X			X	X	X	X	X			X	X	2339	0	2339	2339
Robino *et al.* [48]														X						X	X		X		X	X	X	X	X					X	409	0	409	409
Wang *et al.* [49]	X	X		X	X	X	X	X		X	X	X	X	X	X	X	X		X	X		X			X	X			X	X			X	X	981	0	981	981
Antão-Sousa *et al.* [50]	X	X		X	X	X	X	X		X	X	X	X		X	X														X			X		1598	0	1598	1598
Adnan *et al.* [51]	X	X		X	X	X	X	X		X	X	X	X		X	X														X			X		418	0	0	0

Reference	Markers																																Number of father-son duos used to:			
	DYS19 [1,2]	DYS385a/b* [1,2]	DYS388	DYS389I/II [1,2]	DYS390 [1,2]	DYS391 [1,2]	DYS392 [1,2]	DYS393 [1,2]	DYS435	DYS437 [1,2]	DYS438 [1,2]	DYS439 [1,2]	DYS448 [1,2]	DYS449 [2]	DYS456 [1,2]	DYS458 [1,2]	DYS460 [2]	DYS461	DYS481 [1,2]	DYS518 [2]	DYS526a/b*	DYS533 [1,2]	DYS547	DYS549 [1]	DYS570 [1,2]	DYS576 [1,2]	DYS612	DYS626	DYS627 [2]	DYS635 (C4) [1,2]	DYS643 [1]	GATA A10	GATA H4 [1,2]	DYF387S1 ** [2]	Average mutation rates analyses (Results 3.1.)	Age analyses (Results 3.2.)	Alelle length analyses (Results 3.3.)	
																																				Repeats gain or loss	Single or multistep mutations
Bugoye *et al.* [52]	X	X		X	X	X	X	X		X	X	X	X		X	X						X							X			X		100	0	0	0
Mertoglu *et al.* [53]	X	X		X	X	X	X	X		X	X	X	X		X	X						X							X			X		90	0	0	0
Yang *et al.* [54]	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X						X							X			X		578	0	0	0
Wu *et al.* [55]	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X		X	X		X			X	X			X	X		X	X	1160	0	0	0
Petrovic *et al.* [56]	X	X		X	X	X	X	X		X	X	X	X		X	X			X			X		X	X	X				X	X		X		269	0	0	0
Zhang *et al.* [57]	X	X		X	X	X	X	X		X	X	X	X	X	X	X	X		X	X		X			X	X			X	X			X	X	101	0	101	101
Ayse *et al.* [58]														X						X	X	X			X	X	X	X	X					X	100	0	0	0
Yuan *et al.* [63]														X						X	X	X			X	X	X	X	X					X	501	0	0	0
Ambrosio *et al.* [59]	X	X		X	X	X	X	X		X	X	X	X		X	X			X			X		X	X	X				X	X		X		407	390	407	407
Lin *et al.* [60]	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	640	0	0	0
Fu *et al.* [61]	X	X		X	X	X	X	X		X	X	X	X		X	X	X		X			X		X	X	X				X	X		X		175	0	0	0
Bredemeyer *et al.* [62]	X	X		X	X	X	X	X		X	X	X	X	X	X	X	X		X		X	X	X	X	X	X	X	X	X	X	X		X	X	64	0	0	0
GHEP collaborative exercise	X	X		X	X	X	X	X		X	X	X	X	X	X	X	X		X	X				X	X	X			X	X	X		X	X	806 to 2225	1935	806 to 2225	806 to 2225

Included in the PowerPlex® Y23 System[1], or Yfiler™ Plus PCR Amplification kit[2]

*Multi-copy marker with the same number of loci among individuals

**Multi-copy marker with a variable number of loci among individuals

Table S2: Number of mutations and allele transmissions observed considering data obtained from literature (see Table S1) and from this study, excepting marker DYF387S1 (Table S3).

Markers	Previous works		This work				
	Nº of mutations	Nº of allele transmissions (Total)	Nº of mutations	Nº of allele transmissions (Total)	Paternal allele	Filial allele	Number of observed mutations
DYS19	53	24 148	4	2 224	14	15	3
					15	16	1
DYS385a/b	107	42 217	5	4 442	13	11	1
					13	14	1
					15	16	2
					20	19	1
DYS388	2	3 612	0	0	-	-	-
DYS389I	61	23 929	8	2 225	12	13	1
					13	12	1
					13	14	3
					14	13	1
					14	15	1
					15	14	1
DYS389II-I	83	23 898	8	2 215	16	17	1
					17	16	1
					17	18	3
					18	17	1
					18	19	1
					19	18	1
DYS390	57	22 879	6	2 224	24	23	2
					24	25	2
					25	24	1
					26	25	1
DYS391	62	23 567	4	2 222	10	11	2
					11	10	1
					11	12	1
DYS392	14	22 044	0	2 222	-	-	-
DYS393	29	22 113	2	2 219	12	13	1
					13	12	1
DYS435	0	161	0	0	-	-	-
DYS437	21	18 793	1	2 225	16	15	1
DYS438	9	18 808	1	2 225	11	12	1
DYS439	94	18 887	16	2 224	10	11	1
					11	10	1
					11	12	3
					12	11	3
					12	13	3
					13	12	2
					13	14	1

Markers	Previous works		This work				
	Nº of mutations	Nº of allele transmissions (Total)	Nº of mutations	Nº of allele transmissions (Total)	Paternal allele	Filial allele	Number of observed mutations
					14	13	1
					15	14	1
DYS448	19	14 357	1	2 220	20	19	1
DYS449	87	7 175	12	1 305	26	27	1
					29	30	1
					30	29	1
					30	31	2
					31	30	2
					31	32	1
					32	31	2
					32	33	1
					33	32	1
DYS456	65	15 667	16	2 225	14	15	1
					15	14	1
					15	16	3
					16	15	5
					16	17	1
					17	16	3
					17	18	2
DYS458	110	15 695	16	2 225	14	12	1
					14	13	1
					15	16	3
					16	17	3
					17	16	1
					17	18	2
					17.2	18.2	1
					18	17	1
					20	19	2
					21	20	1
DYS460	19	4 297	4	1 308	10	11	1
					11	10	1
					12	11	2
DYS461	0	873	0	0	-	-	-
DYS481	15	3 390	10	2 114	22	21	1
					22	23	3
					23	24	1
					24	23	1
					25	24	1
					26	27	1
					29	28	1
					30	29	1
DYS518	97	6 492	19	1 303	37	36	1
					37	38	1
					38	36	1

Markers	Previous works		This work				
	Nº of mutations	Nº of allele transmissions (Total)	Nº of mutations	Nº of allele transmissions (Total)	Paternal allele	Filial allele	Number of observed mutations
					38	37	1
					38	39	1
					39	38	1
					39	40	4
					40	39	2
					41	40	3
					41	42	1
					42	43	1
					43	44	1
					46	45	1
DYS526a	12	4 118	0	0	-	-	-
DYS526b	48	4 117	0	0	-	-	-
DYS533	8	4 863	1	2 112	13	12	1
DYS547	69	4 053	0	0	-	-	-
DYS549	7	1 811	3	806	12	11	1
					13	12	1
					14	13	1
DYS570	73	7 862	14	2 114	15	16	1
					16	17	4
					17	16	2
					17	18	3
					18	17	1
					18	19	1
					19	18	1
					24	23	1
DYS576	104	7 763	36	2 113	16	15	1
					16	17	2
					17	16	1
					18	16	1
					18	17	7
					18	19	3
					19	18	7
					19	20	3
					20	19	8
					20	21	1
					21	20	1
					22	21	1
DYS612	66	4 056	0	0	-	-	-
DYS626	41	4 441	0	0	-	-	-
DYS627	111	6 723	22	1 305	19	20	1
					20	19	3
					20	21	2
					21	20	1
					21	22	2

Markers	Previous works		This work				
	Nº of mutations	Nº of allele transmissions (Total)	Nº of mutations	Nº of allele transmissions (Total)	Paternal allele	Filial allele	Number of observed mutations
					22	21	2
					22	23	4
					23	22	2
					23	24	2
					24	25	2
					25	24	1
DYS635	50	16 589	9	2 224	21	20	2
					22	21	3
					23	22	2
					23	24	1
					25	26	1
DYS643	1	1 172	0	806	-	-	-
GATA A10	4	1 026	0	0	-	-	-
GATA H4	39	15 388	3	2 223	11	10	1
					12	11	1
					12	13	1
Overall	1 637	416 984	221	53 070			

Table S3: Observed number of mutations considering data obtained from literature (Table S1) and those generated in this study for marker DYF387S1.

Previous works		This work						
Nº of mutations	Nº of duos	Nº of mutations	Nº of duos	Paternal haplotype		Filial haplotype		Number of observed mutations
66	6746	12	1308	35	36	36	36	1
				35	37	36	37	1
				35	38	36	38	1
				36	38	36	39	1
				37	37	36	37	1
				37	39	36	39	1
				37	38	37	39	1
				37	40	37	41	1
				38	39	38	38	1
				39	40	39	39	1
				39	41	39	40	2

Table S4: Number of transmissions and mutations according to father's age (in years). Novel data added to those produced in two previous GHEP-ISFG working commissions [29, 39] and from [59].

Paternal age (years old)	Number of allele transmissions	Number of mutations observed
9	23	0
14	50	0
15	201	1
16	457	0
17	771	5
18	1527	4
19	2008	10
20	2958	11
21	3251	9
22	3316	9
23	4235	9
24	3906	11
25	4374	15
26	4402	12
27	4110	11
28	4253	7
29	3709	14
30	3852	15
31	3899	12
32	3755	11
33	3196	12
34	2832	8
35	2960	10
36	2279	16
37	1765	6
38	2236	9
39	1396	7
40	1677	5
41	1243	3
42	1080	6
43	980	2
44	815	6
45	862	3
46	680	5
47	731	5

Paternal age (years old)	Number of allele transmissions	Number of mutations observed
48	608	3
49	589	7
50	488	5
51	387	3
52	299	1
53	420	3
54	244	1
55	179	0
56	196	2
57	240	1
58	185	2
59	127	0
60	130	1
61	154	1
62	102	0
63	63	0
64	147	1
65	34	0
66	58	1
67	53	0
68	50	1
69	17	0
70	17	0
71	60	0
73	26	0
74	12	0
79	41	0
Total	84 715	302

Table S5: P-values for the pairwise comparisions between classes of parental age, considering the occurrence, or not, of at least one mutation in the corresponding genetic transmission. Significant values (α=0.05) are showed in red, and those nearly significant in blue. For absolute values see Table 2 in the main text.

		Paternal age (years old)					
		<21	21-30	31-40	41-50	51-60	>61
Paternal age (years old)	<21	-	-	-	-	-	-
	21-30	0,146055	-	-	-	-	-
	31-40	0,852203	0,064179425	-	-	-	-
	41-50	0,117419	0,000146494	0,02492	-	-	-
	51-60	0,20383	0,01171581	0,116791	0,888165	-	-
	>61	0,687829	0,314424003	0,621164	0,772994	0,732614	-

	<21	>20
<21	-	-
>20	0,663765	-

	<31	>30
<31	-	-
>30	0,003020863	-

	<41	>40
<41	-	-
>40	0,000175085	-

	<51	>50
<51	-	-
>50	0,058303357	-

		<61	>60
	<61	-	-
	>60	0,562278549	-

Mutation matrixes

Mutation matrix 1: mutation matrix for marker DYS19. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	11	12	13	14	14.1	14.2	14.3	15	16	17	18	19	15, 16	14, 15, 17	Total
11	91	0	0	0	0	0	0	0	0	0	0	0	0	0	91
12	0	13	1	0	0	0	0	0	0	0	0	0	0	0	14
13	0	0	1357	1	0	0	0	0	0	0	0	0	0	0	1358
14	0	0	1	5323	0	1	0	9	0	0	0	0	0	0	5334
14.1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
14.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14.3	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2
15	0	0	0	0	0	0	0	2991	3	0	0	0	0	0	2994
16	0	0	0	0	0	0	0	2	1296	2	0	0	0	0	1300
17	0	0	0	0	0	0	0	0	5	580	2	0	0	0	587
18	0	0	0	0	0	0	0	0	0	0	5	0	0	0	5
19	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
15, 16	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
14, 15, 17	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Total															11689

Mutation matrix 2: mutation matrix for marker DYS388. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.
s

	10	11	12	13	14	Total
10	17	0	1	0	0	18
11	0	1	0	0	0	1
12	0	0	73	0	0	73
13	0	0	0	7	0	7
14	0	0	0	0	2	2
Total						101

Mutation matrix 3: mutation matrix for marker DYS389I. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	8	9	10	11	12	13	14	15	16	17	18	Total
8	1	0	0	0	0	0	0	0	0	0	0	1
9	0	71	0	0	0	0	0	0	0	0	0	71
10	0	0	57	0	0	0	0	0	0	0	0	57
11	0	0	0	85	0	0	0	0	0	0	0	85
12	0	0	0	1	2206	5	0	0	0	0	0	2212
13	0	0	0	0	4	5997	10	0	0	0	0	6011
14	0	0	0	0	0	12	2084	2	0	0	0	2098
15	0	0	0	0	0	0	3	63	0	0	0	66
16	0	0	0	0	0	0	0	0	9	0	0	9
17	0	0	0	0	0	0	0	0	0	3	0	3
18	0	0	0	0	0	0	0	0	0	0	3	3
Total												10616

Mutation matrix 4: mutation matrix for marker DYS389II-I. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	10	11	12	13	14	15	16	17	18	19	20	16,17	16,18	Total
10	1	0	0	0	0	0	0	0	0	0	0	0	0	1
11	0	8	0	0	0	0	0	0	0	0	0	0	0	8
12	0	0	14	0	0	0	0	0	0	0	0	0	0	14
13	0	0	0	51	1	0	0	0	0	0	0	0	0	52
14	0	0	0	0	98	1	0	0	0	0	0	0	0	99
15	0	0	0	0	0	710	2	0	0	0	0	0	0	712
16	0	0	0	0	0	0	5446	9	1	0	0	0	0	5456
17	0	0	0	0	0	1	8	3068	8	0	0	0	0	3085
18	0	0	0	0	0	0	0	9	979	3	0	0	0	991
19	0	0	0	0	0	0	0	0	4	156	0	0	0	160
20	0	0	0	0	0	0	0	0	0	1	11	0	0	12
16,17	0	0	0	0	0	0	0	0	0	0	0	2	0	2
16,18	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Total														10593

Mutation matrix 5: mutation matrix for marker DYS390. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	18	19	20	21	22	23	24	25	26	27	Total
18	1	0	0	0	0	0	0	0	0	0	1
19	0	1	0	0	0	0	0	0	0	0	1
20	0	0	9	0	0	0	0	0	0	0	9
21	0	0	0	483	2	0	0	0	0	0	485
22	0	0	0	0	1123	0	0	0	0	0	1123
23	0	0	0	0	1	3022	2	0	0	0	3025
24	0	0	0	0	0	8	5159	7	0	0	5174
25	0	0	0	0	0	0	5	1812	0	0	1817
26	0	0	0	0	0	0	0	4	175	0	179
27	0	0	0	0	0	0	0	0	0	12	12
Total											11826

Mutation matrix 6: mutation matrix for marker DYS391. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	5	6	8	9	10	11	12	13	14	Total
5	1	0	0	0	0	0	0	0	0	1
6	0	7	0	0	0	0	0	0	0	7
8	0	0	15	0	0	0	0	0	0	15
9	0	0	0	608	1	0	0	0	0	609
10	0	0	0	1	6765	9	0	0	0	6775
11	0	0	0	0	12	4234	10	0	0	4256
12	0	0	0	0	0	2	148	0	0	150
13	0	0	0	0	0	0	0	13	0	13
14	0	0	0	0	0	0	0	0	1	1
Total										11827

Mutation matrix 7: mutation matrix for marker DYS392. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	7	9	10	10.2	11	11.1	12	13	14	15	16	17	Total
7	2	0	0	0	0	0	0	0	0	0	0	0	2
9	0	10	0	0	0	0	0	0	0	0	0	0	10
10	0	0	71	0	0	0	0	0	0	0	0	0	71
10.2	0	0	0	2	0	0	0	0	0	0	0	0	2
11	0	0	1	0	4649	0	0	0	0	0	0	0	4650
11.1	0	0	0	0	0	1	0	0	0	0	0	0	1
12	0	0	0	0	0	0	801	0	0	0	0	0	801
13	0	0	0	0	0	0	0	4791	3	0	0	0	4794
14	0	0	1	0	0	0	0	0	1296	2	0	0	1299
15	0	0	0	0	0	0	0	0	0	150	0	0	150
16	0	0	0	0	0	0	0	0	0	0	27	0	27
17	0	0	0	0	0	0	0	0	0	0	0	3	3
Total													11810

Mutation matrix 8: mutation matrix for marker DYS393. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	9	10	11	12	12.3	13	14	15	16	Total
9	1	0	0	0	0	0	0	0	0	1
10	0	19	0	0	0	0	0	0	0	19
11	0	0	55	0	0	0	0	0	0	55
12	0	0	0	2053	0	2	0	0	0	2055
12.3	0	0	0	0	1	0	0	0	0	1
13	0	0	0	4	0	6840	4	0	0	6848
14	0	0	0	0	0	1	1327	2	0	1330
15	0	0	0	0	0	0	2	242	0	244
16	0	0	0	0	0	0	0	2	22	24
Total										10577

Mutation matrix 9: mutation matrix for marker DYS437. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	12	13	13.1	13.2	14	15	16	17	18	15, 17	Total
12	1	0	0	0	0	0	0	0	0	0	1
13	0	33	0	0	0	0	0	0	0	0	33
13.1	0	0	2	0	0	0	0	0	0	0	2
13.2	0	0	0	1	0	0	0	0	0	0	1
14	0	0	0	0	4717	3	0	0	0	0	4720
15	0	0	0	0	6	4249	3	0	0	0	4258
16	0	0	0	0	0	1	1011	1	0	0	1013
17	0	0	0	0	0	0	1	26	0	0	27
18	0	0	0	0	0	0	0	0	1	0	1
15, 17	0	0	0	0	0	0	0	0	0	1	1
Total											10057

Mutation matrix 10: mutation matrix for marker DYS438. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	6	7	8	9	9.2	10	11	11.2	12	13	14	Total
6	0	0	0	0	0	0	0	0	0	0	0	0
7	0	6	0	0	0	0	0	0	0	0	0	6
8	0	0	31	0	0	0	0	0	0	0	0	31
9	0	0	0	833	0	0	0	0	0	0	0	833
9.2	0	0	0	0	2	0	0	0	0	0	0	2
10	1	0	0	2	0	3422	0	0	0	0	0	3425
11	0	0	0	0	0	0	2017	0	1	0	0	2018
11.2	0	0	0	0	0	0	0	25	0	0	0	25
12	0	0	0	0	0	2	0	0	3433	0	0	3435
13	0	0	0	0	0	0	0	0	0	330	0	330
14	0	0	0	0	0	0	0	0	0	0	14	14
Total												10119

Mutation matrix 11: mutation matrix for marker DYS439. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	8	9	10	11	11.1	12	13	14	15	10,11	11,11	11,12	Total
8	5	0	0	0	0	0	0	0	0	0	0	0	5
9	0	30	0	0	0	0	0	0	0	0	0	0	30
10	0	0	1025	2	0	0	0	0	0	0	0	0	1027
11	0	0	3	3257	1	7	0	0	0	0	0	0	3268
11.1	0	0	0	0	1	0	0	0	0	0	0	0	1
12	0	0	0	6	0	4240	12	0	0	0	0	0	4258
13	0	0	0	0	0	13	1268	9	0	0	0	0	1290
14	0	0	0	0	0	0	8	180	0	0	0	0	188
15	0	0	0	0	0	0	0	2	6	0	0	0	8
10, 11	0	0	0	0	0	0	0	0	0	3	0	0	3
11, 11	0	0	0	0	0	0	0	0	0	0	0	1	1
11, 12	0	0	0	0	0	0	0	0	0	0	0	0	0
Total													10079

Mutation matrix 12: mutation matrix for marker DYS448. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	15	16	17	18	18.2	19	19.2	20	21	22	23	24	25	19,20	Total
15	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2
16	0	19	0	0	0	0	0	0	0	0	0	0	0	0	19
17	0	0	71	0	0	0	0	0	0	0	0	0	0	0	71
18	0	0	0	982	0	1	0	0	0	0	0	0	0	0	983
18.2	0	0	0	0	8	0	0	0	0	0	0	0	0	0	8
19	0	0	0	4	0	3430	0	2	0	0	0	0	0	0	3436
19.2	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2
20	0	0	0	0	0	3	0	3010	0	0	0	0	0	0	3013
21	0	0	0	0	0	1	0	1	1032	0	0	0	0	0	1034
22	0	0	0	0	0	0	0	0	1	160	0	0	0	0	161
23	0	0	0	0	0	0	0	0	0	0	20	0	0	0	20
24	0	0	0	0	0	0	0	0	0	0	0	5	0	0	5
25	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
19,20	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Total															8756

Mutation matrix 13: mutation matrix for marker DYS449. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	24	25	26	27	27.2	28	28.2	29	29.2	29.3	30	30.2	31	31.2	32	32.2	33	34	35	36	37	38	40	Total
24	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
25	0	49	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	49
26	0	0	144	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	145
27	0	0	0	190	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	192
27.2	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
28	0	0	0	0	0	462	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	462
28.2	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
29	0	0	0	0	0	4	0	900	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	909
29.2	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
29.3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
30	0	0	0	0	0	0	0	8	0	0	941	0	3	0	0	0	0	0	0	0	0	0	0	952
30.2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
31	0	0	0	0	0	0	0	0	0	0	5	0	938	0	5	0	0	0	0	0	0	0	0	948
31.2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	2
32	0	0	0	0	0	0	0	0	0	0	0	0	5	0	821	0	8	0	0	0	0	0	0	834
32.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	3
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	565	4	0	0	0	0	0	573
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	288	1	0	0	0	0	295
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	98	3	0	0	0	104
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	48	0	0	0	49
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	5
38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	2
40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Total																								5538

Mutation matrix 14: mutation matrix for marker DYS456. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	10	11	12	13	14	15	15.3	16	17	18	19	20	15,16	Total
10	1	0	0	0	0	0	0	0	0	0	0	0	0	1
11	0	3	0	0	0	0	0	0	0	0	0	0	0	3
12	0	0	25	0	0	0	0	0	0	0	0	0	0	25
13	0	0	0	210	0	0	0	0	0	0	0	0	0	210
14	0	0	0	1	971	2	0	0	0	0	0	0	0	974
15	0	0	0	0	2	3991	0	12	0	0	0	0	0	4005
15.3	0	0	0	0	0	0	6	0	0	0	0	0	0	6
16	0	0	0	0	0	14	0	2448	4	0	0	0	0	2466
17	0	0	0	0	1	0	0	7	870	7	0	0	0	885
18	0	0	0	0	0	0	0	0	3	143	0	0	0	146
19	0	0	0	0	0	0	0	0	0	0	33	0	0	33
20	0	0	0	0	0	0	0	0	0	0	0	1	0	1
15,16	0	0	0	0	0	0	0	0	0	0	0	0	2	2
Total														8757

Mutation matrix 15: mutation matrix for marker DYS460. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	7	8	9	10	11	12	13	Total
7	5	0	0	0	0	0	0	5
8	0	5	0	0	0	0	0	5
9	0	0	437	1	0	0	0	438
10	0	0	1	1345	3	0	0	1349
11	0	0	0	4	1461	1	0	1466
12	0	0	0	0	4	137	0	141
13	0	0	0	0	0	0	7	7
Total								3411

Mutation matrix 16: mutation matrix for marker DYS481. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	Total
17	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
18	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
19	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	15
20	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	27
21	0	0	0	0	232	0	0	0	0	0	0	0	0	0	0	0	232
22	0	0	0	0	1	1216	4	0	0	0	0	0	0	0	0	0	1221
23	0	0	0	0	0	0	766	3	0	0	0	0	0	0	0	0	769
24	0	0	0	0	0	0	1	463	3	0	0	0	0	0	0	0	467
25	0	0	0	0	0	0	1	1	438	0	0	0	0	0	0	0	440
26	0	0	0	0	0	0	0	0	0	234	3	0	0	0	0	0	237
27	0	0	0	0	0	0	0	0	0	0	177	0	0	0	0	0	177
28	0	0	0	0	0	0	0	0	0	0	1	83	1	0	0	0	85
29	0	0	0	0	0	0	0	0	0	0	0	1	22	0	0	0	23
30	0	0	0	0	0	0	0	0	0	0	0	0	1	8	0	0	9
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Total																	3707

Mutation matrix 17: mutation matrix for marker DYS458. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	12	13	14	14.1	15	15.1	15.2	16	16.1	16.2	17	17.2	18	18.2	19	19.2	20	20.2	21	21.1	21.2	22	17.2, 17.2	17.2, 18.2	Total
12	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
13	0	101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	101
14	1	2	268	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	271
14.1	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
15	0	0	2	0	1330	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1337
15.1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
15.2	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
16	0	0	0	0	3	0	0	2122	0	0	9	0	0	0	1	0	0	0	0	0	0	0	0	0	2135
16.1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
16.2	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21
17	0	0	0	0	1	0	0	11	0	0	2596	0	12	0	0	0	0	0	0	0	0	0	0	0	2620
17.2	0	0	0	0	0	0	0	0	0	0	0	62	0	0	0	0	0	0	0	0	0	0	0	0	62
18	0	0	0	0	0	0	0	1	0	0	7	0	1399	0	6	0	0	0	0	0	0	0	0	0	1413
18.2	0	0	0	0	0	0	0	0	0	0	0	0	0	68	0	0	0	0	0	0	0	0	0	0	68
19	0	0	0	0	0	0	0	0	0	0	0	0	3	0	476	0	6	0	0	0	0	0	0	0	485
19.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33	0	0	0	0	0	0	0	0	33
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	150	0	0	0	0	0	0	0	155
20.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	12
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	17	0	0	1	0	0	20
21.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
21.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	9
17.2,17.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
17.2,18.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total																									8765

Mutation matrix 15: mutation matrix for marker DYS518. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	32	33	33.1	34	34.1	35	36	36.2	36.3	37	37.2	37.3	38	38.2	39	39.2	40	40.2	41	42	43	43.3	44	45	46	47	Total
32	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
33	0	21	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22
33.1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
34	0	0	0	61	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	61
34.1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
35	0	0	0	0	0	192	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	194
36	0	0	0	0	0	1	400	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	402
36.2	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
36.3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
37	0	0	0	0	0	0	3	0	0	718	0	0	6	0	0	0	1	0	0	0	0	0	0	0	0	0	728
37.2	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
37.3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
38	0	0	0	0	0	0	1	0	0	6	0	0	1151	0	6	0	0	0	1	1	0	0	0	0	0	0	1166
38.2	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	4
39	0	0	0	0	0	0	0	0	0	0	0	0	9	0	1032	0	8	0	0	0	0	0	0	0	0	0	1049
39.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2
40	0	0	0	0	0	0	0	0	0	0	0	0	1	0	7	0	794	0	3	1	0	0	0	0	0	0	806
40.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
41	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	7	0	497	7	0	0	0	0	0	0	512
42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	292	8	0	0	0	0	0	302
43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	139	0	4	0	0	0	145
43.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	48	0	1	0	50
45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	21	0	0	22
46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	6	0	7
47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Total																											5494

Mutation matrix 16: mutation matrix for marker DYS526a. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	10	11	12	13	14	15	16	17	18	19	20	Total
10	1	0	0	0	0	0	0	0	0	0	0	1
11	0	98	0	0	0	0	0	0	0	0	0	98
12	0	0	301	1	0	0	0	0	0	0	0	302
13	0	1	2	484	2	0	0	0	0	0	0	489
14	0	0	1	1	996	1	1	0	0	0	0	1000
15	0	0	0	0	2	904	0	1	0	0	0	907
16	0	0	0	0	0	0	272	1	0	0	0	273
17	0	0	0	0	0	0	0	44	0	0	0	44
18	0	0	0	0	0	0	0	0	3	0	0	3
19	0	0	0	0	0	0	0	0	0	1	0	1
20	0	0	0	0	0	0	0	0	0	0	1	1
Total												3119

Mutation matrix 20: mutation matrix for marker DYS533. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	6	7	8	9	10	11	12	12.1	13	13.1	14	15	16.3	Total
6	1	0	0	0	0	0	0	0	0	0	0	0	0	1
7	0	1	0	0	0	0	0	0	0	0	0	0	0	1
8	0	0	6	0	0	0	0	0	0	0	0	0	0	6
9	0	0	0	57	0	0	0	0	0	0	0	0	0	57
10	0	0	0	0	261	0	0	0	0	0	0	0	0	261
11	0	0	0	0	0	1391	1	0	0	0	0	0	0	1392
12	0	0	0	0	0	1	1712	0	1	0	0	0	0	1714
12.1	0	0	0	0	0	0	0	2	0	0	0	0	0	2
13	0	0	0	0	0	0	2	0	219	0	0	0	0	221
13.1	0	0	0	0	0	0	0	0	0	1	0	0	0	1
14	0	0	0	0	0	0	0	0	0	0	44	0	0	44
15	0	0	0	0	0	0	0	0	0	0	0	1	0	1
16.3	0	0	0	0	0	0	0	0	0	0	0	0	4	4
Total														3705

Mutation matrix 21: mutation matrix for marker DYS549. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	8	9	10	11	12	13	14	15	Total
8	1	0	0	0	0	0	0	0	1
9	0	1	0	0	0	0	0	0	1
10	0	0	30	0	0	0	0	0	30
11	0	0	0	208	0	0	0	0	208
12	0	0	0	2	488	1	0	0	491
13	0	0	0	0	3	477	0	0	480
14	0	0	0	0	0	1	91	0	92
15	0	0	0	0	0	0	0	14	14
Total									1317

Mutation matrix 22: mutation matrix for marker DYS526b. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	15	17	18	19	20	21	22	22.2	23	24	24.1	25	26	27	32	33	34	35	36	37	38	20,25	Total
15	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
17	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
18	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16
19	0	0	0	64	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	65
20	0	0	0	0	300	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	300
21	0	0	0	0	2	452	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	457
22	0	0	0	0	0	1	765	0	6	1	0	0	0	0	0	0	0	0	0	0	0	0	773
22.2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
23	0	0	0	0	0	0	6	0	758	5	0	0	0	0	0	0	0	0	0	0	0	0	769
24	0	0	0	0	0	0	0	0	4	471	0	7	0	0	0	0	0	0	0	0	0	0	482
24.1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
25	0	0	0	0	0	0	0	0	0	0	0	172	0	0	0	0	0	0	0	0	0	0	172
26	0	0	0	0	0	0	0	0	0	0	0	2	31	0	0	0	0	0	0	0	0	0	33
27	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	6
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	2
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	7
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	2
38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
20,25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Total																							3095

Mutation matrix 23: mutation matrix for marker DYS547. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	40	41	42	43	44	45	46	46.2	47	47.2	48	48.2	49	49.2	49.3	50	50.2	51	52	53	54	55	Total
40	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
41	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
42	0	2	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
43	0	0	0	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	50
44	0	0	0	0	92	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	93
45	0	0	0	0	0	199	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	200
46	0	0	0	0	0	1	329	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	335
46.2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
47	0	0	0	0	0	0	5	0	507	0	5	0	0	0	0	0	0	0	0	0	0	0	517
47.2	0	0	0	0	0	0	0	0	0	11	0	1	0	0	0	0	0	0	0	0	0	0	12
48	0	0	0	0	0	0	0	0	4	0	769	0	5	0	0	0	0	0	0	0	0	0	778
48.2	0	0	0	0	0	0	0	0	0	0	0	99	0	0	0	0	0	0	0	0	0	0	99
49	0	0	0	0	0	0	0	0	0	0	6	0	603	0	0	2	0	0	0	0	0	0	611
49.2	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	19
49.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
50	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	360	0	2	0	0	0	0	372
50.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	4
51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	97	1	0	0	0	101
52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	36	0	0	0	38
53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	9	0	0	10
54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2
55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Total																							3259

Mutation matrix 24: mutation matrix for marker DYS570. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	12	13	13.3	14	15	16	17	17.2	18	19	19.3	20	20.2	20.3	21	21.2	22	23	24	25	Total
12	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
13	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7
13.3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
14	0	0	0	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36
15	0	0	0	0	111	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	113
16	0	0	0	0	0	736	5	0	0	0	0	0	0	0	0	0	0	0	0	0	741
17	0	0	0	1	1	4	2042	0	10	2	0	0	0	0	0	0	0	0	0	0	2060
17.2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
18	0	0	0	0	0	0	8	0	1709	6	0	0	0	0	0	0	0	0	0	0	1723
19	0	0	0	0	0	0	0	0	6	1332	0	4	0	0	0	0	0	0	0	0	1342
19.3	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	5
20	0	0	0	0	0	0	0	0	2	3	0	601	0	0	3	0	0	0	0	0	609
20.2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2
20.3	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	2
21	0	0	0	0	0	0	0	0	0	0	0	4	0	0	190	0	0	0	0	0	194
21.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	4
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	114	1	0	0	117
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	38	0	0	38
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	15	0	16
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3
Total																					7015

Mutation matrix 25: mutation matrix for marker DYS576. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	12	12.1	13	14	15	16	17	17.3	18	18.2	19	19.1	20	20.2	21	22	23	24	17, 18	Total
12	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
12.1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
13	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16
14	0	0	0	66	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	66
15	0	0	0	0	291	1	0	0	0	0	0	0	0	0	0	0	0	0	0	292
16	0	0	0	0	1	988	6	0	0	0	0	0	0	0	0	0	0	0	0	995
17	0	0	0	0	0	2	1727	0	10	0	0	0	0	0	0	0	0	0	0	1739
17.3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
18	0	0	0	0	0	1	18	0	2156	0	15	0	0	0	0	0	0	0	0	2190
18.2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
19	0	0	0	0	0	2	0	0	19	0	1119	0	12	0	0	0	0	0	0	1152
19.1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
20	0	0	0	0	0	0	0	0	1	0	9	0	363	0	4	0	0	0	0	377
20.2	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	4
21	0	0	0	0	0	0	0	0	0	0	0	0	5	0	64	0	0	0	0	69
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	6	0	0	0	7
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
17, 18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18, 18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Total																				6914

Mutation matrix 26: mutation matrix for marker DYS612. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	25	27	28	30	31	32	33	34	35	36	37	38	39	40	40.1	41	41.1	42	42.1	44	36, 38	38, 38	Total
25	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
27	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
28	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
30	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
31	0	0	0	0	9	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10
32	0	0	0	0	0	37	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40
33	0	0	0	0	0	0	82	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	84
34	0	0	0	0	0	1	0	257	1	0	0	0	0	0	0	0	0	0	0	0	0	0	259
35	0	0	0	0	0	0	0	4	470	8	1	0	0	0	0	0	0	0	0	0	0	0	483
36	0	0	0	0	0	0	0	0	3	927	3	0	0	0	0	0	0	0	0	0	0	0	933
37	0	0	0	0	0	0	0	0	0	6	697	7	0	0	0	0	0	0	0	0	0	0	710
38	0	0	0	0	0	0	0	0	0	1	5	420	5	1	0	0	0	0	0	0	0	0	432
39	0	0	0	0	0	0	0	0	0	0	0	2	220	1	0	0	0	0	0	0	0	0	223
40	0	0	0	0	0	0	0	0	0	0	0	0	0	46	0	0	0	0	0	0	0	0	46
40.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33	0	0	0	0	0	0	0	33
41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	17
41.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	6
42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	2
42.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36, 38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38, 38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
Total																							3288

Mutation matrix 27: mutation matrix for marker DYS626. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	22	23	24	25	26	27	28	29	29.1	30	30.2	31	32	33	33.2	34	35	36	39	45.2	Total
22	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
23	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
24	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18
25	0	0	0	115	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	115
26	0	0	0	0	177	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	177
27	0	0	0	1	1	148	0	0	0	0	0	0	0	0	0	0	0	0	0	0	150
28	0	0	0	0	0	0	188	2	0	0	0	0	0	0	0	0	0	0	0	0	190
29	0	0	0	0	0	0	1	585	0	1	0	0	0	0	0	0	0	0	0	0	587
29.1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
30	0	0	0	0	0	0	0	1	0	608	0	1	0	0	0	0	0	0	0	0	610
30.2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
31	0	0	0	0	0	0	0	0	0	5	0	589	2	0	0	0	0	0	0	0	596
32	0	0	0	0	0	0	0	0	0	0	0	4	428	3	0	0	0	0	0	0	435
33	0	0	0	0	0	1	0	0	0	0	0	0	6	238	0	1	0	0	0	0	246
33.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
34	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	84	1	0	0	0	87
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	15	0	0	0	16
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	2
39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
45.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Total																					3237

Mutation matrix 28: mutation matrix for marker DYS627. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	13	14	15	16	16.2	17	17.1	17.2	18	18.2	19	19.2	20	20.1	20.2	20.3	21	21.2	22	22.2	23	23.3	24	25	26	27	Total
13	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
14	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
15	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25
16	0	0	0	294	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	295
16.2	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
17	0	0	0	0	0	431	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	431
17.1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
17.2	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
18	0	0	0	0	0	1	0	0	407	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	411
18.2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
19	0	0	0	0	0	1	0	0	2	0	638	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	644
19.2	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
20	0	0	0	0	0	0	0	0	1	0	5	0	965	0	0	0	10	0	1	0	0	0	0	0	0	0	982
20.1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2
20.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	3
20.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
21	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	1137	0	6	0	0	0	0	0	0	0	1152
21.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
22	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	13	0	950	0	7	0	0	0	0	0	971
22.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	2
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	498	0	5	0	0	0	512
23.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	152	4	0	0	159
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	38	1	0	41
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	11	0	12
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Total																											5663

Mutation matrix 29: mutation matrix for marker DYS635. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	17	18	19	20	21	22	23	24	25	26	27	Total
17	8	0	0	0	0	0	0	0	0	0	0	8
18	0	3	0	0	0	0	0	0	0	0	0	3
19	0	0	175	0	0	0	0	0	0	0	0	175
20	0	0	0	773	0	0	0	0	0	0	0	773
21	0	0	0	4	2245	3	0	0	0	0	0	2252
22	0	0	0	0	12	1200	1	0	0	0	0	1213
23	0	0	0	0	0	6	4132	3	0	0	0	4141
24	0	0	0	0	0	0	2	873	1	0	0	876
25	0	0	0	0	0	0	0	1	171	2	0	174
26	0	0	0	0	0	0	0	0	0	19	0	19
27	0	0	0	0	0	0	0	0	0	0	1	1
Total												9635

Mutation matrix 30: mutation matrix for marker GATA A10. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	13	14	15	16	17	18	Total
13	33	0	0	0	0	0	33
14	1	281	1	0	0	0	283
15	0	2	431	0	0	0	433
16	0	0	0	110	0	0	110
17	0	0	0	0	14	0	14
18	0	0	0	0	0	1	1
Total							874

Mutation matrix 31: mutation matrix for marker GATA H4. The first column represents the paternal (origin) alleles, and the first line represents the filial (final) alleles.

	8	9	10	11	12	13	14	15	16	12, 13	Total
8	2	0	0	0	0	0	0	0	0	0	2
9	0	39	0	0	0	0	0	0	0	0	39
10	0	0	420	1	0	0	0	0	0	0	421
11	0	0	2	3422	3	0	0	0	0	0	3427
12	0	0	0	9	5160	4	0	0	0	0	5173
13	0	0	0	0	2	732	1	0	0	0	735
14	0	0	0	0	0	0	46	0	0	0	46
15	0	0	0	0	0	0	0	2	0	0	2
16	0	0	0	0	0	0	0	0	0	0	0
12, 13	0	0	0	0	0	0	0	0	0	3	3
Total											9848

2.1.4. X-Chromosomal STRs: metapopulations and mutation rates

L. Gusmão[1,#], S. Antão-Sousa[2,3,#], M. A. Abovich[4], D. Aguirre[5], R. Alghafri[6], C. Alves[2], A. Amorim[2,3], C. Arévalo[7], L. Baldassarri[8], C. Barletta-Carrillo[9], G. Berardi[10], C. Bobillo[11], L. Borjas[12], A. Brehm[13], J. J. Builes[5], L. Cainé[14], M. Carvalho[15], E. F. Carvalho[1], L. Catelli[16], R. M. B. Cicarelli[17], D. Corach[11], M. V. Diederiche[18], P. Domingues[1], M. Espinoza[19], M. Faustino[2,3], D. F. Braganholi[17], J. M. Fernandéz[20], O. García[21], M. G. García[22], F. G. Di Marco[22], A. Gaviria[23], I. Gomes[2], D. Grattapaglia[24], J. Henao[25], A. Hernandez[26], A. A. Ibarra[27], G. Lima[14], I. M. Manterola[28], C. Marrero[29], M. M. Pancorbo[30], J. A. Martins[31], L. Mendoza[5], A. Mosquera[32], A. Contreras[33], C. Onofri[34], J. J. Pestano[35], G. Plaza Teixidó[36], M. L. Rebelo[14], M. J. Porto[15], Y. C. Posada[27], E. Riego[37], D. Sumita[38], R. Rodenbusch[39], A. Rodríguez[19], A. Rodríguez[32], P. Sanchez-Diz[40], S. Santos[41], F. Simão[1], L. M. Siza Fuentes[42], E. C. Nascimento[43], C. Tomas[44], U. Toscanini[10], A. Trindade Filho[45], C. Turchi[34], C. Vullo[16], I. Yurrebaso[21], V. Pereira[44,#], N. Pinto[2,46,#]

[1] DNA Diagnostic Laboratory, State University of Rio de Janeiro (UERJ), Rio de Janeiro, Brazil
[2] Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP) and Instituto de Investigação e Inovação em Saúde (i3S), Porto, Portugal
[3] Departamento de Biologia, Faculdade de Ciências da Universidade do Porto (FCUP), Porto, Portugal
[4] Banco Nacional de Datos Genéticos, Buenos Aires, Argentina and Sección Histocompatibilidad, Unidad Inmunología e Histocompatibilidad, Hospital General de Agudos Dr. Carlos G. Durand, CABA, Buenos Aires, Argentina
[5] Laboratorio Genes SAS, Medellín, Colombia
[6] General Department of Forensic Sciences and Criminology, Dubai Police General Head Quarters, Dubai, United Arab Emirates
[7] Laboratorio Biología-ADN, Comisaría General de Policía Científica, Madrid, Spain and Instituto Universitario de Investigación en Ciencias Policiales (IUICP), Universidad de Alcalá de Henares, Madrid, Spain
[8] Laboratorio di Genetica Forense de la Università Cattolica del Sacro Cuore di Roma, Rome, Italy
[9] Laboratorio de Genética Humana, Universidad Nacional Mayor de San Marcos, Lima, Peru
[10] Laboratorio de Inmunogenética y ADN de la Fundación Favaloro (PRICAI), Buenos Aires, Argentina
[11] Servicio de Huellas Digitales Genéticos (SHDG) and Cátedra de Genética y Bioquímica Molecular, Facultad de Farmacia y Bioquímica, Universidad de Buenos Aires, Argentina
[12] Laboratorio de Genética Molecular, Unidad de Genética Médica, Facultad de Medicina, Universidad del Zulia, Zulia, Venezuela
[13] Laboratório de Genética Humana, Universidade da Madeira, Campus da Penteada, Funchal, Portugal
[14] Serviço de Genética e Biologia Forense, Instituto Nacional de Medicina Legal e Ciências Forenses, I.P. – Delegação do Norte, Porto, Portugal
[15] Serviço de Genética e Biologia Forense, Instituto Nacional de Medicina Legal e Ciências Forenses, I.P. – Delegação do Centro, Coimbra, Portugal
[16] DNA Forensic Laboratory, Argentinean Forensic Anthropology Team (EAAF), Córdoba, Argentina
[17] Laboratório de Investigação de Paternidade-NAC, Faculdade de Ciências Farmacêuticas, Universidade Estadual Paulista (UNESP), Araraquara, São Paulo, Brazil
[18] Departamento de Biologia da Universidade Estadual de Santa Cruz, Salvador, Brazil
[19]Departamento de Ciencias Forenses, Sección de Bioquímica, Unidad de Genética Forense, Poder Judicial, San José, Costa Rica
[20] Departamento de Biología, Servicio de Criminalística, Dirección General de la Policía y la Guardia Civil, ámbito Guardia Civil, Spain

[21] Sección de Genética Forense, Area de Laboratorio Ertzaintza, Bizkaia, Spain
[22] Laboratorio ManLab, Area de Filiaciones, Buenos Aires, Argentina
[23] Laboratorio de Genética Molecular and Hemocentro Nacional - Cruz Roja Ecuatoriana, Quito, Ecuador
[24] Heréditas Tecnologia em Análise de DNA, Brasília, Brazil
[25] Laboratorio de Genética Médica, Universidad Tecnológica de Pereira, Facultad de Ciencias de la Salud, Pereira, Colombia
[26] Instituto Nacional de Toxicología y Ciencias Forenses, Delegación de Canarias, Santa Cruz de Tenerife, Spain
[27] Laboratorio IdentiGEN - Universidad de Antioquia, Medellín, Colombia
[28] Servicio Genómica - SGIker - Universidad del País Vasco (UPV-EHU), Bilbao, Spain
[29] Laboratorio Genomik C.A., Valencia, Venezuela
[30] Banco de ADN, Universidad del País Vasco (UPV/EHU), Vitoria, Gasteiz, Spain
[31] Research Centre for Biochemistry and Molecular Biology at the Medical School of São José do Rio Preto (FAMERP), São José do Rio Preto, São Paulo, Brazil
[32] Forensic Genetics Unit, University of Santiago de Compostela, Santiago de Compostela, Spain
[33] Laboratorio Regional de Genética Forense - Poder Judicial de Rio Negro, Rio Negro, Argentina
[34] Institute of Legal Medicine, Università Politecnica delle Marche, Ospedali Riuniti, Ancona, Italy
[35] Laboratorio de Genética Forense, Facultad de Medicina, Universidad de Las Palmas de Gran Canaria, Las Palmas, Spain
[36] NEODIAGNOSTICA, SL, Lleida, Spain
[37] Unidad de Parentesco e Identificación Humana por ADN, Referencia Laboratorio Clínico, Dominican Republic
[38] Genomic Engenharia Molecular Ltda., São Paulo, Brasil
[39] Fundação Estadual de Produção e Pesquisa em Saúde (FEPPS) – Laboratório de Investigação de Paternidade (LIP), Porto Alegre, Brazil
[40] NASERTIC, Navarra, Spain
[41] Human and Medical Genetics Laboratory, Federal University of Pará, Belém, Brazil
[42] Genetica Molecular de Colombia Ltda, Bogota, Colombia
[43] Coordenação de Genética Forense, Departamento de Polícia Técnica da Bahia, Salvador, Brazil
[44] Section of Forensic Genetics - Department of Forensic Medicine, Faculty of Health and Medical Sciences – University of Copenhagen, Copenhagen, Denmark
[45] Instituto de Pesquisa de DNA Forense – Polícia Civil do Distrito Federal, Brasília, Brasil
[46] Centro de Matemática da Universidade do Porto (CMUP), Porto, Portugal

[#] These authors have contributed equally to this work

In final preparation. For supplementary material please check Appendix C.

Keywords: Forensic genetics; Kinship analysis; Population stratification; Mutation rates; X-STRs

1. Introduction

In the last two decades, studies have been performed to investigate the relevance of using X-chromosomal specific STRs (X-STRs) in forensic genetics (Szibor *et al.*, 2003; Diegoli, 2015; Gomes *et al.*, 2020). The presence of a single X chromosome in males, which is only transmitted to the daughters, increases the informative power of X versus autosomal loci in some specific kinship investigations (Szibor *et al.*, 2003). However, if the case requires the identification of a male individual the X-chromosomal markers have lower discrimination power than the autosomal ones. Furthermore, in male/female mixtures, the presence of up to two alleles in females and only one allele in males decreases the chance of retrieving a male profile, despite increasing the probability of recovering female alleles (Diegoli, 2015).

Thus, X-chromosomal markers are essentially useful in the analysis of biological kinship, especially in complex situations of identification through relatives, where autosomal markers are not able to produce conclusive results. In the simplest case of paternity with access to the presumed father, when compared to equally diverse autosomal STRs, the X-STRs have higher mean exclusion chance (Szibor *et al.*, 2003). Thus, X-STRs have been shown to be useful as a complementary tool in paternity cases where few genetic inconsistencies are observed for autosomal STRs. Similarly, these markers are very useful to solve paternity cases in which there are two questioned fathers that are related as father and son. In this context, the two men will not have the same X chromosome but will transmit it to the daughter. Another advantage of the use of X-chromosomal markers in kinship analysis is in situations where the exclusion power is null for autosomal markers, e.g., in a paternity investigation with only access to the paternal grandmother. In this case, there is an allele that is necessarily transmitted to the granddaughter (through the father), and the absence of this sharing indicates an exclusion, only explained by mutation. The same applies to cases in which the paternal sisterhood or half-sisterhood is investigated, since the father necessarily transmits his only X chromosome to all the daughters. Other examples where X-STRs surpass the informative power of autosomal STRs to resolve kinship issues can be found in several publications (Pinto *et al.*, 2011; Diegoli, 2015; Tillmar *et al.*, 2017; Pereira, Gusmão, 2021).

The recognition of the usefulness of X-chromosomal markers applied to complex kinship analyses prompted the development of genotyping methodologies for X-STRs, and the screening of allele/haplotype frequency distributions in many populations worldwide. Alongside the development of genotyping multiplexes based on PCR-CE methodologies

(e.g. Bini *et al.*, 2005; Asamura *et al.*, 2006; Turrina *et al.*, 2007; Gusmão *et al.*, 2009; Ribeiro-Rodrigues *et al.*, 2007; Edelmann *et al.*, 2012; Liu *et al.*, 2012; Prieto-Fernández *et al.*, 2016; Zhang *et al.*, 2021), X-STRs have also been incorporated into forensic marker panels designed for MPS (Jäger *et al.*, 2017). In parallel, interpretation software was developed to account for the specific characteristics associated to X-chromosomal loci, namely the specific mode of inheritance, mutation, linkage, and linkage disequilibrium between loci (Kling *et al.*, 2015).

For any genetic marker to be used in forensic casework, it is crucial to know the genetic profile of the relevant populations. The evaluation of forensic genetic evidence is based on likelihood ratio (LR) principles, considering the probability of the observed profile(s) under two alternative and mutually exclusive hypotheses (Gjertson *et al.*, 2007). The probability is usually calculated based on population allele frequencies, being, however, more complex when it comes to the combined analysis of markers in linkage disequilibrium (LD). Since marker independence can no longer be assumed, the final LR cannot be obtained by multiplying the individual values obtained for each locus. The most direct implication for the use of markers in LD in kinship analyses is the need to estimate haplotype frequencies in the population, rather than single locus allele frequencies (Tillmar *et al.*, 2017).

The underlying difficulties in estimating haplotype frequencies are notorious for non-recombinant markers such as the Y chromosome and mtDNA, where different approaches have been proposed to calculate the probability of the evidence based on these markers (Egeland, Salas, 2008; Andersen *et al.*, 2020; Andersen, Balding, 2021). Although to a lesser extent, the difficulty on estimating haplotype frequencies is also applied to X-STRs that are in LD, since the expected number of allelic combinations usually exceeds the hundreds (Kling, 2018). Therefore, very large population samples are needed for a proper estimation of the haplotype frequencies in a population.

Another important parameter to consider in the application of genetic markers in kinship analyses is the mutation rate. Due to the low frequency and multiple factors associated with the probability of mutation in a given allele, the amount of data needed to obtain reliable estimates is very high. The studies carried out so far on X-STRs show a behavior similar to that of autosomal STRs. X-chromosomal marker mutation rates present great variation between and within STRs, that depends on the number and structure of the repetitive motif, and the gender and age of the parents (Pinto *et al.*, 2020).

In summary, the usefulness of X-STRs is well recognized in kinship analysis. There are well-established genotyping and result interpretation methodologies. However, although there are many reports on X-STR allele/haplotype frequencies in several populations,

the information available is still fragmentary both in terms of the populations analyzed and the markers included (Gomes *et al.*, 2020). Moreover, the mutational data available is still scarce to encompass the variation that is known to exist between and within X-STRs.

Therefore, this study aimed to generate and compile data on X-STRs in different populations worldwide, as well as data on segregation analyses in father/mother/daughter trios. These data comprised 19 X-STRs widely used in forensic genetics, which have been described in two PCR multiplexes with three overlapping loci: one with 10 X-STRs (X-Decaplex) and the other with 12 X-STRs (Argus X-12).

A joint analysis of the data obtained with those available in the literature allowed for the comparison of the different populations, and the establishment of diversity and substructure levels. Additionally, more thorough mutation rates were estimated based on segregation analyses on father/mother/daughter trios.

2. Materials and Methods

2.1. Sampling and genotyping methods

Data for this study was obtained through collaborative works of the Spanish and Portuguese Working Group of the International Society for Forensic Genetics (GHEP-ISFG). A total of 42 laboratories have participated by producing data on 1,395 father/mother/daughter trios genotyped for the X-Decaplex, and 878 trios genotyped for the Investigator Argus X-12 QS (Qiagen) loci. Only the daughters were used for population analyses, after inferring the gametic phase using the information from the father. Samples from 6066 unrelated males typed for the X-Decaplex were also included. Samples belonged to paternity cases or to healthy volunteers. In samples from paternity investigations, the biological relationship was previously confirmed using autosomal STRs (LR > 10,000). Each laboratory ensured the anonymization of the samples and the accomplishment of the legal and ethical requirements for their use in this research project.

The 10 loci included in the X-Decaplex were genotyped as described in Gusmão *et al.* (2009). In the genotyping with the kit Investigator Argus X-12 QS (Qiagen), the manufacturer's instructions were followed (Investigator® Argus X-12 QS Handbook).

2.2. Population data and analyses

Population analyses were performed after joining data from the present study and previously published data, whenever information for X-Decaplex haplotypes or Argus Linkage Groups (LG1 to LG4) was available. The final dataset included 187 population samples from 53 countries in Africa, America, Europe, and Asia (see Fig. 1).

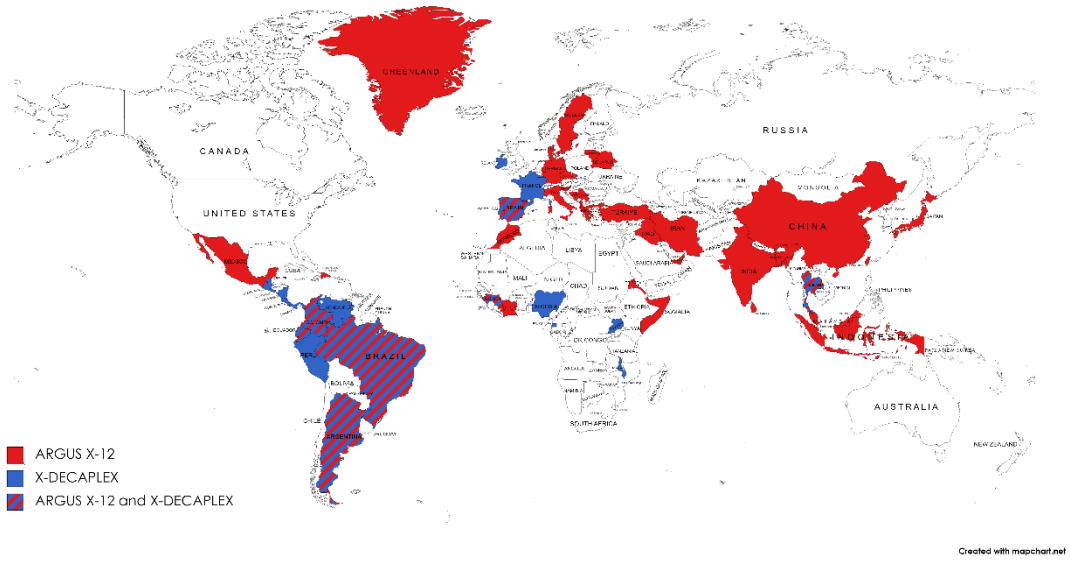


Figure 1. Map of the populations from which data was collected for X-DECAPLEX, ARGUS-X12 or both marker sets.

For the X-Decaplex markers, a total of 12,440 haplotypes were compiled (Toscanini *et al.* 2009; Valente *et al.* 2009; Gomes *et al.* 2009; Martins *et al.* 2010; Pereira *et al.* 2012; Glesmann *et al.* 2013; Lopez *et al.* 2015; Prieto-Fernandez *et al.* 2016; Prieto-Fernandez *et al.* 2017; Martins *et al.* 2017; Rincon *et al.* 2017; Baeta *et al.* 2021), including 8,911 new haplotypes from this study (Supplementary Table S1). From the literature, we have only included data from complete male X-Decaplex profiles. Data from Gusmão *et al.* (2009) was not considered, since the male profiles in that article correspond to fathers of the daughter included in the present study. The gametic phase of the daughters was inferred using the information from the father.

For the Argus X-12, a total of 11,509 full haplotypes were collected (Pasino *et al.* 2011; Tillmar *et al.* 2011; Edelmann *et al.* 2012; Tomas *et al.* 2012, Uchigasaki *et al.* 2013; Samejima *et al.* 2012; Zidkova *et al.* 2014; Xiao-na *et al.* 2014; Pinto *et al.* 2015; Pereira *et al.* 2015; Rębała *et al.* 2015; Tomas *et al.* 2015; Poulsen *et al.* 2015; Shrivastava *et al.* 2015; Poulsen *et al.* 2016; Vongpaisarnsin *et al.* 2016; Gomes *et al.* 2017; Crnjac *et al.* 2017, Mrsic *et al.* 2017; Ferragut *et al.* 2017; Guo *et al.* 2017; Sufian *et al.* 2017; Mrsic

et al. 2018, Robino *et al.* 2018; Veselinović *et al.* 2018; Almarri *et al.* 2018; Tao *et al.* 2018; Salvador *et al.* 2018; Cortés-Trujillo *et al.* 2019; Ferragut *et al.* 2019; Xing *et al.* 2019; Li *et al.* 2019; Garcia *et al.* 2019; Kalaoglu *et al.* 2019; Pinto *et al.* 2020; Hering *et al.* 2020; Xiao *et al.* 2020; Kakkar *et al.* 2020; Bini *et al.* 2021; Ferragut *et al.* 2021; Flores *et al.* 2021; Perera *et al.* 2021a; Perera *et al.* 2021b; Hakim *et al.* 2021; Khacha-ananda *et al.* 2021), 1,533 new haplotypes from this study. From the literature, we have only included data from male profiles with LGs' haplotypic information. Some publications report the frequencies of LG haplotypes, but do not include the full 12 X-STR haplotypes. Therefore, sample size varied from 17,814 for LG1 and LG2; 17,349 for LG3; and 16,732 for LG4 (Supplementary Table S1).

Allele and haplotype frequencies were calculated using the software Arlequin ver 3.5.2.2 (Excoffier and Lischer, 2010). The same software was used for genetic structure analyses, based on population pairwise comparisons using F_{ST} values and corresponding non-differentiation p-values. Pairwise linkage disequilibrium was also tested using Arlequin ver 3.5.2.2 (Excoffier and Lischer, 2010). In population differentiation and LD tests, Bonferroni's correction was used to adjust the significance level, by dividing 0.05 by the total number of pairwise comparisons.

2.3. Segregation data and mutation analyses

Data from 3,748 father/mother/daughter trios was compiled from the literature and joined to the 2,273 trios from this study to estimate X-STR mutation rates (Supplementary Table S2). Segregation data from father/daughter or mother/son duos was not considered in the present study, since the frequency of hidden mutations increases in these cases, when compared to father/mother/daughter trios (Antão-Sousa *et al.*, 2019).

Confidence intervals for mutation rates were estimated from the binomial standard deviation.

3. Results and discussion

3.1. Genetic structure analyses

With the aim of investigating the different levels of genetic differentiation within and between countries and continents for the studied X-STRs, pairwise comparisons were performed between available population samples. Since the set of samples studied for the X-Decaplex and for the Argus X-12 loci do not overlap, analyses were performed separately for each of these sets and the results interpreted based on the overall

findings. Populations from the same country or continental region were grouped, unless they showed large (≥0.0100) and statistically significant F_{STs} for the studied markers.

3.1.1. Pairwise population comparisons based on the X-Decaplex data

A hierarchical analysis was performed per country or continent, depending on the number of representative populations.

Pairwise comparisons, showed non-significant F_{ST} values between populations from the West-Central African region (Supplementary Table S3). Although not statistically significant, the F_{ST} between Guinea-Bissau and Equatorial Guinea was relatively high, which can be due to the small sample size of the latter. Malawi, the only population from Southeast Africa, presented low F_{STs} with all populations from the West-Central African region, except Guinea-Bissau. As expected, both Morocco, in North Africa, and the Eastern population of a Nilotic speaking group from the Uganda (Karimojong) presented significant differences between them and with the remaining populations, except with Equatorial Guinea, most likely for the reasons mentioned above.

Concerning the two Asian populations from Macau and Thailand, a low, non-significant, F_{ST} was observed (F_{ST}= 0.0006; p=0.40253).

When comparing populations from Central America, the Native American groups from Guatemala and El Salvador showed non-significant differences (Supplementary Table S4; Supplementary Figure S1). In pairwise F_{ST} analysis among the admixed populations, all values were below 1%, except between Guatemala and Panama. Between these two countries, the non-differentiation probability was not significant, although the F_{ST} was relatively high (0.01416). In contrast, a significant differentiation was observed between Costa Rica and Nicaragua, but for a low F_{ST} value, most probably due to the large sample sizes.

Among the 10 samples from Argentina, the three Native groups showed significant differences among them and with the remaining samples from non-Native populations (Supplementary Table S5). Except in comparisons involving the Northeastern populations of Corrientes and Misiones, all pairwise F_{STs} were low (F_{ST} ≤0.0053), with no statistically significant differentiation detected among the populations from the Central and South regions, for the studied markers.

The 14 samples from Brazil were analyzed considering the main geographic regions of the country: Southeast (including Araraquara, Belo Horizonte, São Paulo, Rio de Janeiro and Vitoria); Northeast (Bahia); North (Belem), Central west (Brasilia, Mato Grosso, Mato Grosso do Sul); and South (Parana, Santa Catarina and Rio Grande do Sul). Although

statistically significant differences were observed in some pairwise comparisons of population from the same region, the F_{STs} were all below 1% (Supplementary Table S6; Supplementary Figure S2). The same was observed between samples from Southeast, Central west, and South regions. However, statistically significant F_{STs}, above 1%, were observed in pairwise comparison involving populations from the North and Northeast regions.

In the comparison of the Colombian samples, high F_{ST} values were observed between Choco and the other populations (Supplementary Table S7; Supplementary Figure S3), which is somehow expected due to the high African ancestry reported for this region in the Pacific Coast (Ossa *et al.*, 2016). The same trend is observed for Bolivar, although for smaller distances. Lower F_{STs} were observed among the remaining populations, all located in the Andean region. However, all pairwise comparisons involving Bogota resulted in F_{ST} slightly above 1% ($0.0102 \leq F_{ST} \leq 0.0110$) preventing the grouping of this population with the remaining ones from the Andes.

The three samples from Ecuador presented large, statistically significant F_{STs}, as expected considering that they correspond to different ethnic groups, namely the Kichwa, the Waorani and an admixed population sample (known as Mestizos). The F_{ST} values were 0.04407 between Kichua and Mestizos ($P<0.000005$), 0.06778 between Waorani and Mestizos ($P<0.000005$), and 0.05898 between Kichua and Waorani ($P=0.00099$).

Between the two population samples from Venezuela (Central region and Maracaibo), a low, non-significant, F_{ST} was observed ($F_{ST}= 0.0021$; $p=0.0584$).

When comparing European populations within and between countries, only the Portuguese Gypsies and Canarias populations showed significant differences in some pairwise comparisons, with F_{ST} values above 1% (Supplementary Table S8). No significant differences were observed among the remaining populations from Portugal, Italy, Ireland, Spain and France (Supplementary Table S8).

Based on the above-described results, a new pairwise F_{ST} analysis was performed after grouping country/continent neighboring populations with no evidence of presenting large distances ($F_{STs}<0.0100$). The results obtained are displayed in Supplementary Table S9 and represented in the MDS plot in Figure 2. In the MDS, it is possible to identify some population groupings (with low F_{STs}), with population isolates being highly dispersed in the plot, evidencing strong drift or endogamy effects. The sample from South American populations show a wide distribution with varying distances to the African, European and Native American populations.

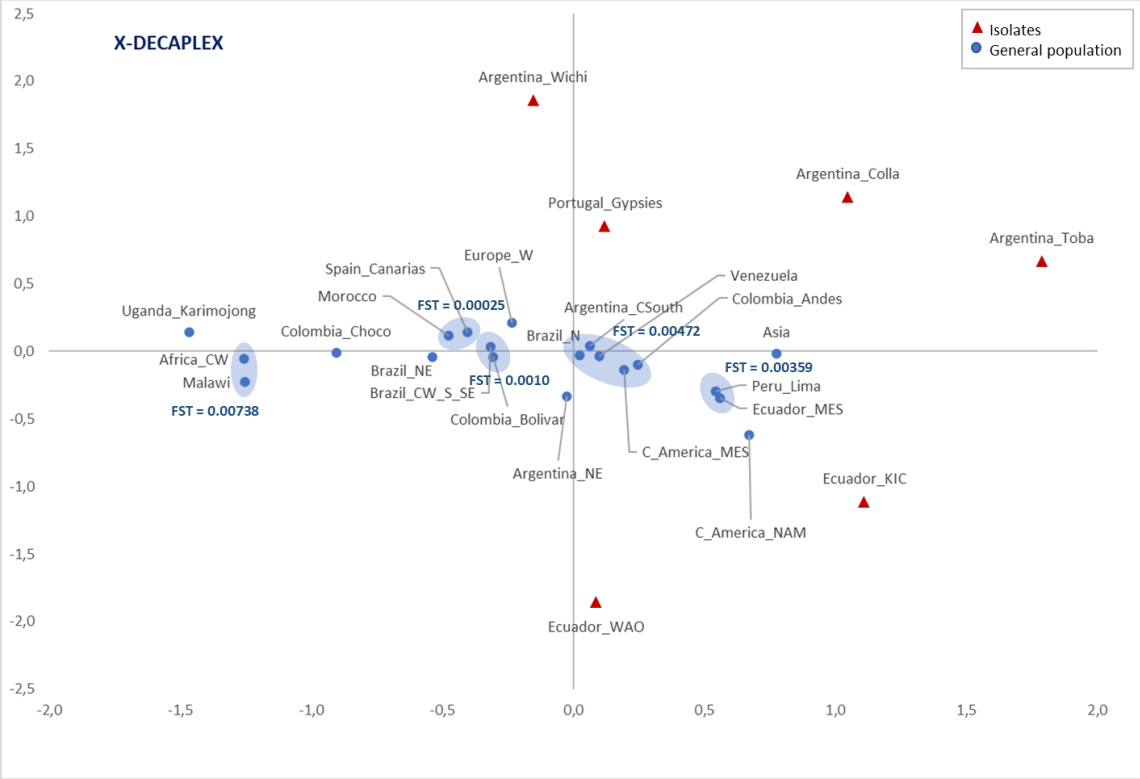


Figure 2. Multidimensional scaling (MDS) plot based on F_{ST} population pairwise genetic distances for the 10 X-STRs included in the X-Decaplex. For details on populations used see Supplementary Table S1. Stress = 0.14934

3.1.2. Pairwise population comparisons based on the Argus X-12 data

As for the X-Decaplex, pairwise F_{ST} analyses at different populational levels were performed, to understand which samples could be group in a single population, assuming a low F_{ST} and non-differentiation probability.

Different populations/ethnic groups reported for the same country were pooled, whenever they were shown not to be significantly different in the original publication. For instance, the samples from the 3 ethnic groups from Morocco were grouped in a single population (called Morocco), since no significant differentiation was reported among them in the original study (Bentayebi *et al.*, 2012). The same was done for the Sri-Lankan ethnicities studied by Perera *et al.* (2021), since low genetic distances were observed among them (F_{STs}≤0.00375). The Zhuang and Mulao from the Chinese province of Guangxi were also joined, based on the results from Xiao *et al.* (2020). In accordance with the findings from Cortés-Trujillo *et al.* (2019), a single population sample from Mexico was considered, including 7 different country regions. Also, in accordance with

the original publication, a single population sample from Iraq was considered, including the Persians, Lurs, Kurds and Azeris ethnic groups (Poulsen *et al.*, 2015). The samples from Sardinian open populations were grouped since no significant differentiation was detected among them, in contrast with the isolate populations that were more heterogeneous (Robino *et al.*, 2018). The nine subpopulations from Greece were pooled in accordance with the results from the original study (Tomas *et al.*, 2015).

Concerning samples from the same population/country/region reported in different studies, a first analysis was performed to see if they were not significantly different. When comparing haplotype frequencies between the different Jewish groups, no significant pairwise F_{STs} were observed for the 4 LGs, and samples were pooled in a single group.

For the 24 populations from Asia, based on the defined threshold of F_{ST} lower than 1% or that were non-statistically significant, and considering the geographic location, populations were compared inside three regions: East, Southeast and South Asia. Although belonging to different ethnic groups (Manchu, Shenyang and Han, from Liaoning, and Korean from Jilin), the 4 population samples from the Northeast region of China were grouped since low, non-significant, pairwise F_{STs} were detected for the 4 LGs (F_{ST} ≤0.0023).

Concerning South America, the results of differentiation analysis between samples from the 23 Argentinean provinces allowed grouping them in 4 regions, namely: Northwest (Catamarca, Jujuy, La Rioja, Salta, Santiago del Estero and Tucuman); Northeast (Chaco, Corrientes, Formosa and Misiones); Central (Buenos Aires, Cordoba, Entre Rios, La Pampa, Mendoza, San Juan, San Luis and Santa Fe); and South (Chubut, Neuquén, Rio Negro, Santa Cruz and Tierra del Fuego). The three Brazilian samples from the Southeast region showed non-significant differences in all LGs (F_{ST} ≤0.0049; P≥0.0276). Statistically non-significant P-values were also obtained when comparing the four Ecuadorian population samples. Nevertheless, high F_{ST} values (between 0.0194 to 0.0259) were observed in some comparisons between the sample from general population described in Pinto *et al.* (2020) and the three population groups included in Flores-Espinosa *et al.* (2021), although not statistically significant (above the significance values of 0.0083), most probably due to the low sample size (n=22).

For the European samples, a first pooling between populations from the same country was performed whenever non-significant F_{STs} were observed. Samples from the 4 regions of Croatia (North, South, East and West) were pooled; all Portuguese samples were treated as a single population; Italy isolated populations were maintained separated from the open populations; and Spain was divided in mainland and Balearic Islands. A second comparison was carried out among all European populations. Based on the F_{ST}

results and corresponding non-differentiation probabilities, the following grouping was performed: East Europe (Belarus, Czech Republic, Lithuania and Slovenia); Southeast (Albania, Croatia, Greece and Serbia); and West (Denmark, Germany, Italy, Portugal, Spain and Sweden).

Based on the above-described results, a new pairwise F_{ST} analysis was performed after grouping country/continent neighboring populations with no evidence of presenting large distances (F_{STs}<0.0100). The results obtained are displayed in Supplementary Table S10 and represented in the MDS plot in Figure 3. In the MDS, we have defined 4 population group, considering the geographic location and the distances among populations. European and Middle Eastern populations (including a group of Jewish populations) cluster together by small distances. Taiwan and Northwest China showed a very low F_{ST}. Brazil and Dominic Republic stand in between the European cluster and the African population from Guinea Bissau, with no significant differentiation between them. The populations from different Argentinean regions, as well as from Mexico could also be group when considering an F_{ST} below 1%. Ecuador separates from the remaining populations from South America (included in the analysis), most probably due to a higher native contribution.

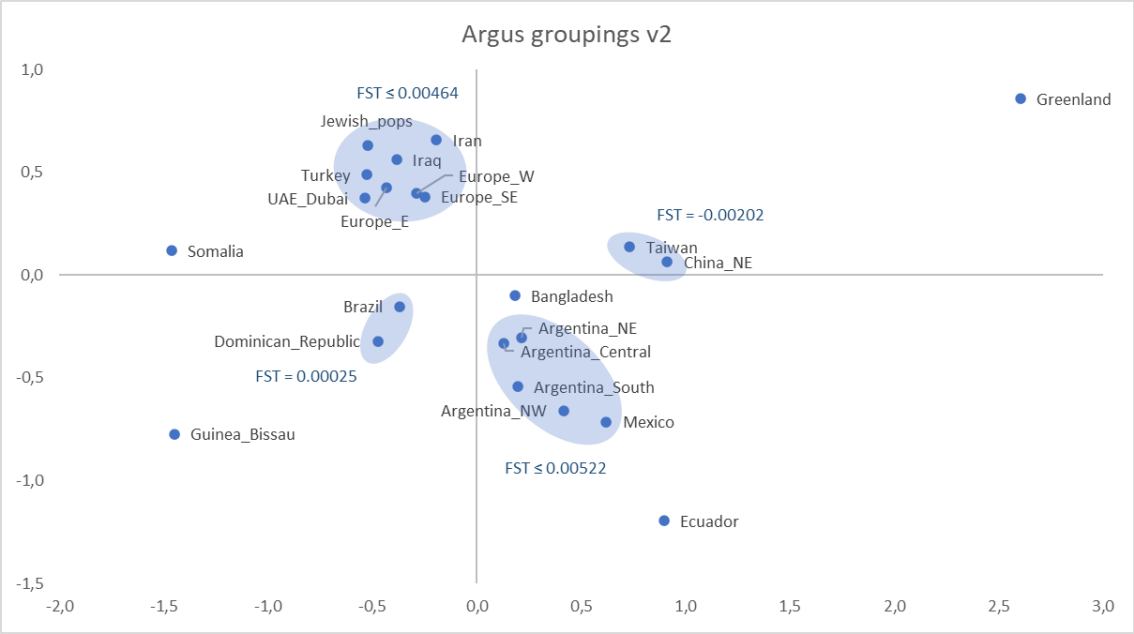


Figure 3. Multidimensional scaling (MDS) plot based on F_{ST} population pairwise genetic distances for the 12 X-STRs included in the Argus-X12. For details on populations used see Supplementary Table S1. Stress = 0.09141

3.2. Segregation and mutation analyses

Among the 2273 father/mother/daughter trios analyzed in this study, 145 incompatibilities were detected (48 out of 1395 trios analyzed for the X-Decaplex, and 97 out of 878 trios analyzed for the Argus X-12). No incompatibilities with simple Mendelian transmission were detected at DXS9898. For the remaining X-STRs, the number of incompatibilities varied from one, at DXS10103 and DXS7133, to 28 at DXS7132 (Supplementary Table S11). Two of the 145 incompatibilities can be explained by null alleles at DXS6789 and GATA172D05, respectively, since both mother and daughter show a single allele at these loci (Supplementary Table S11). The remaining incompatibilities were explained by mutation, with one trio showing both a maternal and a paternal mutation at DXS10101. Among the 143 mutations, 110 were of paternal and 15 of maternal origin, and 18 were indetermined.

3.2.1. Mutation rates

Locus specific mutation rates were calculated after combining our data with published information on father/mother/daughter trios. Data was obtained for 37 X-STRs, including the 19 from this study (Supplementary Table S12). The average mutation rate for the 37 X-STRs was 2.827×10^{-3}, being higher when just considering the ARGUS-X12 (4.021×10^{-3}) than the X-Decaplex markers (1.635×10^{-3}). No mutations were detected in 13 of the loci. In 10 cases, this can be explained by the relatively low number (below 1000) of the studied allelic transmissions, since the upper limit of the 95% confidence interval (CI) is above the average mutation rate for all loci. The results obtained for the other 3 X-STRs (DXS9898, DXS981 and GATA165B12) support low mutation rates. For the remaining loci, mutation rates were variable, with the lowest estimate (0.268×10^{-3} for DXS7133) being 40 times lower than the largest (10.689×10^{-3} for DXS10135).

The mutation rate was 5.2 times higher in the male than in the female germline. This male mutation bias has been frequently described in humans, as well as in other mammals (Wilson Sayres & Makova, 2011; Hurst & Ellegren, 1998), due to the higher number of cell divisions in male than in female germline. Although known to vary with age, the expected male mutation bias in humans was estimated to be around 6 (Hurst & Ellegren, 1998).

References

Almarri, M. A., & Lootah, R. A. (2018). Allelic and haplotype diversity of 12 X-STRs in the United Arab Emirates. *Forensic Science International: Genetics*, *33*, e4-e6.

Andersen MM, Balding DJ. Assessing the Forensic Value of DNA Evidence from Y Chromosomes and Mitogenomes. *Genes* (Basel). 2021 Aug 5;12(8):1209. doi: 10.3390/genes12081209.

Andersen MM, Caliebe A, Kirkeby K, Knudsen M, Vihrs N, Curran JM. Estimation of Y haplotype frequencies with lower order dependencies. *Forensic Sci Int Genet*. 2020 May;46:102214. doi: 10.1016/j.fsigen.2019.102214.

Antão-Sousa, S., Conde-Sousa, E., Gusmão, L., Amorim, A., & Pinto, N. (2019). Underestimation and misclassification of mutations at X chromosome STRs depend on population's allelic profile. *Forensic Science International: Genetics Supplement Series*, *7*(1), 718-720.

Asamura H, Sakai H, Kobayashi K, Ota M, Fukushima H. MiniX-STR multiplex system population study in Japan and application to degraded DNA analysis. *Int J Legal Med*. 2006 May;120(3):174-81. doi: 10.1007/s00414-005-0074-6.

Baeta, M., Prieto-Fernández, E., Núñez, C. *et al.* Study of 17 X-STRs in Native American and Mestizo populations of Central America for forensic and population purposes. *Int J Legal Med* **135**, 1773–1776 (2021). https://doi.org/10.1007/s00414-021-02536-9

Bentayebi, K., Picornell, A., Bouabdeallah, M., Castro, J. A., Aboukhalid, R., Squalli, D., ... & Amzazi, S. (2012). Genetic diversity of 12 X-chromosomal short tandem repeats in the Moroccan population. *Forensic Science International: Genetics*, *6*(1), e48-e49.

Bini C, Ceccardi S, Ferri G, Pelotti S, Alù M, Roncaglia E, Beduschi G, Caenazzo L, Ponzano E, Tasinato P, Turchi C, Buscemi L, Mazzanti M, Tagliabracci A, Toni C, Spinetti I, Domenici R, Presciuttini S. Development of a heptaplex PCR system to analyse X-chromosome STR loci from five Italian population samples. A collaborative study. *Forensic Sci Int*. 2005 Oct 29;153(2-3):231-6. doi: 10.1016/j.forsciint.2005.05.013.

Bini, C., Sarno, S., Tangorra, E., Iuvaro, A., De Fanti, S., Tseghereda, Y. G., ... & Luiselli, D. (2021). 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*(2), 449-453.

Cortés-Trujillo, I., Zuñiga-Chiquette, F., Ramos-González, B., de Loures Chávez-Briones, M., Islas-González, K. L., Betancourt-Guerra, D. A., ... & Rangel-Villalobos, H. (2019). Allele and haplotype frequencies of 12 X-STRs in Mexican population. *Forensic Science International: Genetics*, *38*, e11-e13.

Crnjac, J., Ozretić, P., Merkaš, S., Ratko, M., Lozančić, M., Korolija, M., ... & Mršić, G. (2016). Investigator Argus X-12 study on the population of northern Croatia. *Genetics and Molecular Biology*, *40*, 80-83.

Diegoli TM. Forensic typing of short tandem repeat markers on the X and Y chromosomes. *Forensic Sci Int Genet*. 2015 Sep;18:140-51. doi: 10.1016/j.fsigen.2015.03.013.

Edelmann J, Lutz-Bonengel S, Naue J, Hering S. X-chromosomal haplotype frequencies of four linkage groups using the Investigator Argus X-12 Kit. *Forensic Sci Int Genet*. 2012 Jan;6(1):e24-34. doi: 10.1016/j.fsigen.2011.01.001.

Egeland T, Salas A. Estimating haplotype frequency and coverage of databases. *PLoS One*. 2008;3(12):e3988. doi: 10.1371/journal.pone.0003988.

Excoffier, L., & Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources*, *10*(3), 564-567.

Ferragut, J. F., Bassitta, M., Torrens, V., Albeza, V., Acreche, N., Castro, J. A., ... & Picornell, A. (2019). Analysis of 21 X-chromosome polymorphisms in urban and rural populations in Salta province (north-western Argentina). *International journal of legal medicine*, *133*, 1043-1047.

Ferragut, J. F., Bentayebi, K., Barbaro, A., Ramírez, M., Saguillo, A. Y., Ramon, C., & Picornell, A. (2021). Exploring the Western Mediterranean through X-chromosome. *International Journal of Legal Medicine*, *135*(3), 787-790.

Ferragut, J. F., Bentayebi, K., Pereira, R., Castro, J. A., Amorim, A., Ramon, C., & Picornell, A. (2017). Genetic portrait of Jewish populations based on three sets of X-chromosome markers: Indels, Alu insertions and STRs. *Forensic Science International: Genetics*, *31*, e5-e11.

Flores-Espinoza, R., Paz-Cruz, E., Ruiz-Pozo, V. A., Lopez-Carrera, M., Cabrera-Andrade, A., Gusmão, L., & Burgos, G. (2021). Investigating genetic diversity in admixed populations from Ecuador. *American Journal of Physical Anthropology*, *176*(1), 109-119.

García, M. G., Catanesi, C. I., Penacino, G. A., Gusmão, L., & Pinto, N. (2019). X-chromosome data for 12 STRs: Towards an Argentinian database of forensic haplotype frequencies. *Forensic Science International: Genetics*, *41*, e8-e13.

Gjertson DW, Brenner CH, Baur MP, Carracedo A, Guidet F, Luque JA, Lessig R, Mayr WR, Pascali VL, Prinz M, Schneider PM, Morling N. ISFG: Recommendations on biostatistics in paternity testing. *Forensic Sci Int Genet*. 2007 Dec;1(3-4):223-31. doi: 10.1016/j.fsigen.2007.06.006.

Gomes I, Pinto N, Antão-Sousa S, Gomes V, Gusmão L, Amorim A. Twenty Years Later: A Comprehensive Review of the X Chro-mosome Use in Forensic Genetics. *Front Genet*. 2020 Sep 17; 11:926. doi: 10.3389/fgene.2020.00926.

Gomes, I., Pereira, P. J., 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.

Gomes, I., Pereira, V., Gomes, V., Prata, M. J., Pinto, N., Carracedo, A., ... & Gusmão, L. (2009). The Karimojong from Uganda: genetic characterization using an X-STR decaplex system. *Forensic Science International: Genetics*, *3*(4), e127-e128.

Guo, F. (2017). Population genetic data for 12 X-STR loci in the Northern Han Chinese and StatsX package as tools for population statistics on X-STR. *Forensic Science International: Genetics*, *26*, e1-e8.

Gusmão L, Sánchez-Diz P, Alves C, Gomes I, Zarrabeitia MT, Abovich M, Atmetlla I, Bobillo C, Bravo L, Builes J, Cainé L, Calvo R, Carvalho E, Carvalho M, Cicarelli R, Catelli L, Corach D, Espinoza M, García O, Malaghini M, Martins J, Pinheiro F, João Porto M, Raimondi E, Riancho JA, Rodríguez A, Rodríguez A, Rodríguez Cardozo B, Schneider V, Silva S, Tavares C, Toscanini U, Vullo C, Whittle M, Yurrebaso I, Carracedo A, Amorim A. A GEP-ISFG collaborative study on the optimization of an X-STR decaplex: data on 15 Iberian and Latin American populations. *Int J Legal Med*. 2009 May;123(3):227-34. doi: 10.1007/s00414-008-0309-4.

Hakim, H. M., Khan, H. O., Ismail, S. A., Lalung, J., Kofi, A. E., Aziz, M. Y., ... & Edinur, H. A. (2021). Population data and genetic characteristics of 12 X-STR loci using the Investigator® Argus X-12 Quality Sensor kit for the Kedayan population of Borneo in Malaysia. *International journal of legal medicine*, *135*, 1433-1435.

Hering, S., Klimova, A., & Edelmann, J. (2020). German population data for 18 X-STRs: a hexaplex PCR adding two clusters of X-STRs to the Argus X-12 set and expanding the German haplotype databases. *International Journal of Legal Medicine*, *134*, 2061-2062.

Hurst, L. D., & Ellegren, H. (1998). Sex biases in the mutation rate. *Trends in Genetics*, *14*(11), 446-452.

Jäger AC, Alvarez ML, Davis CP, Guzmán E, Han Y, Way L, Walichiewicz P, Silva D, Pham N, Caves G, Bruand J, Schlesinger F, Pond SJK, Varlaro J, Stephens KM, Holt CL. Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories. *Forensic Sci Int Genet*. 2017 May; 28:52-70. doi: 10.1016/j.fsigen.2017.01.011.

Kakkar, S., Shrivastava, P., Sahajpal, V., Sharma, A., Devi, N., Mandal, S. P., & Kumawat, R. K. (2020). Genetic diversity of X-STR markers in Jat Sikh population of Punjab, India and its comparison with other 39 global populations. *Annals of Human Biology*, *47*(5), 490-497.

Kalaoglu, E., Ay, M., Ulubay, A., Canan, H., Serin, A., & Alper, B. (2019). Population data of 12 X-STRs in Turkey. *Forensic Science International: Genetics Supplement Series*, *7*(1), 632-634.

Khacha-Ananda, S., & Mahawong, P. (2021). Genetic analysis of 12 X-short tandem repeats loci in a northern Thai population. *Medicine, Science and the Law*, *61*(1), 34-43.

Kling D, Dell'Amico B, Tillmar AO. FamLinkX - implementation of a general model for likelihood computations for X-chromosomal marker data. *Forensic Sci Int Genet*. 2015 Jul; 17:1-7. doi: 10.1016/j.fsigen.2015.02.007.

Kling D. Curiosities of X chromosomal markers and haplotypes. Int J Legal Med. 2018 Mar;132(2):361-371. doi: 10.1007/s00414-017-1612-8.

Laura Angela Glesmann, Pablo Francisco Martina, Cecilia Inés Catanesi (2013). Genetic Variation of X-STRs in the Wichí Population from Chaco Province, Argentina. *Human Biology*, 85(5), 687-697

Li, X. N., Zheng, J. L., Yao, J., Ding, M., Pang, H., & Wang, B. J. (2014). Population data of 12 X-chromosome STR loci in Yanbian Korean samples from China. *Forensic Science International: Genetics*, *13*, 143-144.

Li, Y., Zeng, J., Fan, Y., He, G., Xie, M., Gao, T., ... & Wu, J. (2019). Genetic portrait and phylogenetic analysis of an Aksu Uyghur population based on the 19 X-STR system. *International Journal of Legal Medicine*, *133*, 91-93.

Liu QL, Lu DJ, Quan L, Chen YF, Shen M, Zhao H. Development of multiplex PCR system with 15 X-STR loci and genetic analysis in three nationality populations from China. *Electrophoresis*. 2012 Apr;33(8):1299-305. doi: 10.1002/elps.201100558.

Lopez, M., Manrique, A., Aguirre, D., Mendoza, L., Bravo, M.L., Salgar, M., Puerto, Y.R., Gusmão, L., & Builes, J.J. (2015). Genetic data of 10 X-STR in an Afro-descendant population sample of the Department of Chocó—Colombia. *Forensic Science International: Genetics Supplement Series, 5*.

Martins, J.A., Costa, J.C., Paneto, G.G. *et al.* Genetic profile characterization of 10 X-STRs in four populations of the southeastern region of Brazil. *Int J Legal Med* **124**, 427–432 (2010). https://doi.org/10.1007/s00414-010-0478-9

Martins, J.A., Martins, D.P., Oliveira-Brancati, C.I.F. *et al.* Genetic characterization of an X-STR decaplex system in the State of Mato Grosso, Brazil: distribution, forensic efficiency and population structure. *Int J Legal Med* **131**, 1523–1530 (2017). https://doi.org/10.1007/s00414-017-1578-6

Mršić, G., Ozretić, P., Crnjac, J., Merkaš, S., Račić, I., Rožić, S., ... & Korolija, M. (2017). Analysis of 12 X-STR loci in the population of south Croatia. *Molecular biology reports*, *44*, 183-189.

Mršić, G., Ozretić, P., Crnjac, J., Merkaš, S., Sukser, V., Račić, I., ... & Korolija, M. (2018). Expanded Croatian 12 X-STR loci database with an overview of anomalous profiles. *Forensic Science International: Genetics*, *34*, 249-256.

Ossa, H., Aquino, J., Pereira, R., Ibarra, A., Ossa, R. H., Pérez, L. A., ... & Gusmão, L. (2016). Outlining the ancestry landscape of Colombian admixed populations. *PloS one*, *11*(10), e0164414.

Pasino, S., Caratti, S., Del Pero, M., Santovito, A., Torre, C., & Robino, C. (2011). Allele and haplotype diversity of X-chromosomal STRs in Ivory Coast. *International Journal of Legal Medicine*, *125*, 749-752.

Pereira, V, Gusmão, L. The X-Chromosomal STRs in Forensic Genetics: X Chromosome STRs. In Forensic DNA Analysis: Technological Development and Innovative Applications, Pilli, E., Berti, A., *Apple Academic Press Inc*. 2021 pp. 91–112. Hard ISBN: 9781771889056; E-Book ISBN: 9781003043027.

Pereira, V., Gusmão, L., Valente, C.M., Pereira, R.P., Carneiro, J., Gomes, I., Morling, N., Amorim, A., & João Prata, M. (2012). Refining the genetic portrait of Portuguese Roma through X-chromosomal markers. *American journal of physical anthropology, 148 3*, 389-94.

Pereira, V., Tomas, C., Sanchez, J. J., Syndercombe-Court, D., Amorim, A., Gusmao, L., ... & Morling, N. (2015). The peopling of Greenland: further insights from the analysis of genetic diversity using autosomal and X-chromosomal markers. *European Journal of Human Genetics*, *23*(2), 245-251.

Perera, N., Galhena, G., & Ranawaka, G. (2021). Development of a 16 X-STR multiplex PCR system for kinship analysis and its applicability for the Sinhalese population in Sri Lanka. *International Journal of Legal Medicine*, *135*(1), 161-166.

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), 1-12.

Pinto N, Gusmão L, Amorim A. X-chromosome markers in kinship testing: a generalisation of the IBD approach identifying situations where their contribution is crucial. *Forensic Sci Int Genet*. 2011 Jan;5(1):27-32. doi: 10.1016/j.fsigen.2010.01.011.

Pinto N, Pereira V, Tomas C, Loiola S, Carvalho EF, Modesti N, Maxzud M, Marcucci V, Cano H, Cicarelli R, Januario B, Bento A, Brito P, Burgos G, Paz-Cruz E, Díez-Juárez L, Vannelli S, Pontes ML, Berardi G, Furfuro S, Fernandez A, Sumita D, Bobillo C, García MG, Gusmão L. Paternal and maternal mutations in X-STRs: A GHEP-ISFG collaborative study. *Forensic Sci Int Genet*. 2020 May; 46:102258. doi: 10.1016/j.fsigen.2020.102258.

Pinto, J. C., Pereira, V., Marques, S. L., Amorim, A., Alvarez, L., & Prata, M. J. (2015). Mirandese language and genetic differentiation in Iberia: a study using X chromosome markers. *Annals of human biology*, *42*(1), 20-25.

Poulsen, L., Farzad, M. S., Børsting, C., Tomas, C., Pereira, V., & Morling, N. (2015). Population and forensic data for three sets of forensic genetic markers in four ethnic groups from Iran: Persians, Lurs, Kurds and Azeris. *Forensic Science International: Genetics*, *17*, 43-46.

Poulsen, L., Tomas, C., Drobnič, K., Ivanova, V., Mogensen, H. S., Kondili, A., ... & Morling, N. (2016). NGMSElect™ and Investigator® Argus X-12 analysis in population

samples from Albania, Iraq, Lithuania, Slovenia, and Turkey. *Forensic Science International: Genetics*, *22*, 110-112.

Prieto-Fernández E, Baeta M, Núñez C, Zarrabeitia MT, Herrera RJ, Builes JJ, de Pancorbo MM. Development of a new highly effi-cient 17 X-STR multiplex for forensic purposes. *Electrophoresis*. 2016 Jul;37(12):1651-8. doi: 10.1002/elps.201500546.

Prieto-Fernández E, Baeta M, Núñez C, Zarrabeitia MT, Herrera RJ, Builes JJ, de Pancorbo MM. Development of a new highly efficient 17 X-STR multiplex for forensic purposes. *Electrophoresis*. 2016 Jul;37(12):1651-8. doi: 10.1002/elps.201500546. Epub 2016 May 2. PMID: 27060859.

Rębała, K., Kotova, S. A., Rybakova, V. I., Zabauskaya, T. V., Shyla, A. A., Spivak, A. A., ... & Szczerkowska, Z. (2015). Variation of X-chromosomal microsatellites in Belarus within the context of their genetic diversity in Europe. *Forensic Science International: Genetics*, *16*, 105-111.

Ribeiro Rodrigues EM Leite FP, Hutz MH, Palha Tde J, Ribeiro dos Santos AK, dos Santos SE. A multiplex PCR for 11 X chromo-some STR markers and population data from a Brazilian Amazon Region. *Forensic Sci Int Genet*. 2008 Mar;2(2):154-8. doi: 10.1016/j.fsigen.2007.10.179.

Rincón, A.L., Aguirre, D., Mendoza, L., Camacho, J.A., & Builes, J.J. (2017). Genetic data of 10 X-STRs in a population sample of the Tunja city, Department of Boyacá-Colombia. *Forensic Science International: Genetics Supplement Series, 6*.

Robino, C., Lacerenza, D., Aneli, S., Di Gaetano, C., Matullo, G., Robledo, R., & Calò, C. (2018). Allele and haplotype diversity of 12 X-STRs in Sardinia. *Forensic Science International: Genetics*, *33*, e1-e3.

Salvador, J. M., Apaga, D. L. T., Delfin, F. C., Calacal, G. C., Dennis, S. E., & De Ungria, M. C. A. (2018). Filipino DNA variation at 12 X-chromosome short tandem repeat markers. *Forensic Science International: Genetics*, *36*, e8-e12.

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*, *126*, 677-683.

Shrivastava, P., Jain, T., Gupta, U., & Trivedi, V. B. (2015). Genetic polymorphism study on 12 X STR loci of investigator Argus X STR kit in Bhil tribal population of Madhya Pradesh, India. *Legal Medicine*, *17*(3), 214-217.

Sufian, A., Hosen, M. I., Fatema, K., Hossain, T., Hasan, M. M., Mazumder, A. K., & Akhteruzzaman, S. (2017). Genetic diversity study on 12 X-STR loci of investigator® Argus X STR kit in Bangladeshi population. *International journal of legal medicine*, *131*, 963-965.

Szibor R, Krawczak M, Hering S, Edelmann J, Kuhlisch E, Krause D. Use of X-linked markers for forensic purposes. *Int J Legal Med*. 2003 Apr;117(2):67-74. doi: 10.1007/s00414-002-0352-5.

Tao, R., Zhang, J., Bian, Y., Dong, R., Liu, X., Jin, C., ... & 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.

Tillmar AO, Kling D, Butler JM, Parson W, Prinz M, Schneider PM, Egeland T, Gusmão L. DNA Commission of the International Society for Forensic Genetics (ISFG): Guidelines on the use of X-STRs in kinship analysis. Forensic Sci Int Genet. 2017 Jul;29:269-275. doi: 10.1016/j.fsigen.2017.05.005.

Tillmar, A. O. (2012). Population genetic analysis of 12 X-STRs in Swedish population. *Forensic Science International: Genetics*, *6*(2), e80-e81.

Tomas, C., Pereira, V., & Morling, N. (2012). Analysis of 12 X-STRs in greenlanders, danes and somalis using argus X-12. *International journal of legal medicine*, *126*, 121-128.

Tomas, C., Skitsa, I., Steinmeier, E., Poulsen, L., Ampati, A., Børsting, C., & Morling, N. (2015). Results for five sets of forensic genetic markers studied in a Greek population sample. *Forensic Science International: Genetics*, *16*, 132-137.

Toscanini, U., Gusmão, L., Berardi, G., & Raimondi, E. (2009). Genetic data of 10 X-STR in two Native American populations of Argentina. *Forensic Science International: Genetics Supplement Series*, *2*(1), 405-406.

Turrina S, Atzei R, Filippini G, De Leo D. Development and forensic validation of a new multiplex PCR assay with 12 X-chromosomal short tandem repeats. *Forensic Sci Int Genet*. 2007 Jun;1(2):201-4. doi: 10.1016/j.fsigen.2007.01.011. Epub 2007 Feb 26.

Uchigasaki, S., Tie, J., & Takahashi, D. (2013). Genetic analysis of twelve X-chromosomal STRs in Japanese and Chinese populations. *Molecular biology reports*, *40*, 3193-3196.

Valente, C., Gomes, I., Pereira, V., Amorim, A., Gusmão, L., & Prata, M. J. (2009). Association between STRs from the X chromosome in a sample of Portuguese Gypsies. *Forensic Science International: Genetics Supplement Series*, *2*(1), 391-393.

Veselinović, I., Vapa, D., Djan, M., Veličković, N., Veljović, T., & Petrić, G. (2018). Genetic analysis of 12 X-STR loci in the Serbian population from Vojvodina Province. *International journal of legal medicine*, *132*, 405-408.

Vongpaisarnsin, K., Boonlert, A., Rasmeepaisarn, K., & Dangkao, P. (2016). Genetic variation study of 12 X chromosomal STR in central Thailand population. *International journal of legal medicine*, *130*(6), 1497-1499.

Wilson Sayres, M. A., & Makova, K. D. (2011). Genome analyses substantiate male mutation bias in many species. *Bioessays*, *33*(12), 938-945.

Xiao, C., Li, S., Zhang, X., Yang, X., Liu, C., & Chen, L. (2020). Population genetic analysis of Chinese Zhuang and Mulao minorities using AGCU-X19 STR kit. *International Journal of Legal Medicine*, *134*, 501-503.

Xing, J., Adnan, A., Rakha, A., Kasim, K., Noor, A., Xuan, J., ... & Wang, B. (2019). Genetic analysis of 12 X-STRs for forensic purposes in Liaoning Manchu population from China. *Gene*, *683*, 153-158.

Zhang Y, Yu Z, Mo X, Zhao X, Li W, Liu H, Liu C, Wu R, Sun H. Development and validation of a new 18 X-STR typing assay for forensic applications. *Electrophoresis*. 2021 Mar;42(6):766-773. doi: 10.1002/elps.202000168.

Zidkova, A., Capek, P., Horinek, A., & Coufalova, P. (2014). Investigator® Argus X-12 study on the population of Czech Republic: Comparison of linked and unlinked X-STR s for kinship analysis. *Electrophoresis*, *35*(14), 1989-1992.

Appendix C

Gusmão, L. & Antão-Sousa, S., *et al.* (2023). X-Chromosomal STRs: metapopulations and mutation rates

For the supplementary tables and figures please access document: Appendix C_X-Chromosomal STRs: metapopulations and mutation rates at: https://drive.google.com/drive/folders/10uaPx_Bxmh2nJxrYxoSMMd7unp7Q0Mit?usp=share_link

Supplementary Table S1 - Population data for the X-Decaplex and Argus X-12 kits used in this study, and respective references from the literature (data collected until Jun 2021).

Supplementary Table S2 - Number of trios used in the segregation analysis per X-STR. Data were produced in this study or collected from the literature (until Jun 2021).

Supplementary Table S3 - Pairwise comparisons for the African populations considering the X-Decaplex markers. F_{STs} are presented below diagonal and respective p-values are presented above. Significant values after Bonferroni correction ($p<0.00179$) are marked in bold.

Supplementary Table S4 - Pairwise comparisons for the Central American admixed and Native (NAM) populations considering the X-Decaplex markers. F_{STs} are presented below diagonal and respective p-values are presented above. Significant values after Bonferroni correction ($p<0.00238$) are marked in bold.

Supplementary Table S5 - Pairwise comparisons for the Argentinian Native (NAM) and non-Native populations considering the X-Decaplex markers. F_{STs} are presented below diagonal and respective p-values are presented above. Significant values after Bonferroni correction ($p<0.00111$) are marked in bold.

Supplementary Table S6 - Pairwise comparisons for the Brazilian samples considering the X-Decaplex markers. F_{STs} are presented below diagonal and respective p-values are presented above. Significant values after Bonferroni correction ($p<0.00055$) are marked in bold.

Supplementary Table S7 - Pairwise comparisons for the Colombian samples considering the X-Decaplex markers. F_{STs} are presented below diagonal and respective p-values are presented above. Significant values after Bonferroni correction ($p<0.00333$) are marked in bold.

Supplementary Table S8 - Pairwise comparisons for the European populations considering the X-Decaplex markers. F_{STs} are presented below diagonal and respective p-values are presented above. Significant values after Bonferroni correction ($p<0.00020$) are marked in bold.

Supplementary Table S9 - Pairwise comparisons for the X-Decaplex markers after grouping country/continent neighboring populations with no evidence of presenting large distances (FSTs<0.0100). F_{STs} are presented below diagonal and respective p-values are presented above. Significant values after Bonferroni correction are marked in bold.

Supplementary Table S10 - Pairwise comparisons for the Argus-X12 markers after grouping country/continent neighboring populations with no evidence of presenting large distances (FSTs<0.0100). F_{STs} are presented below diagonal and respective p-values are presented above. Significant values after Bonferroni correction are marked in bold.

Supplementary Table S11 - Summary of the mutations detected in this study for markers included in X-Decaplex and Argus X-12 kits, and their most parsimonious parental origin.

Supplementary Table S12 - Locus specific mutation rates calculated after combining data from this study with published information on father/mother/daughter trios. Data was obtained for 37 X-STRs, including the 19 from this study.

Supplementary Figure S1 - Map of the Central American admixed and Native (NAM) populations considered.

Supplementary Figure S2 - MDS of the Argentinian populations considered.

Supplementary Figure S3 - Map of the Brazilian regions considered.

Supplementary Figure S4 - Map of the Colombian regions considered.

2.2. STR mutation rates estimation biases

The methodology most commonly used to study STR mutation rates is the MIA applied to data from genotyping platforms with automated length determination after CE. For autosomal and X-chromosomal STRs mutation estimation, this entails biases as there is an intrinsic uncertainty related not only to the occurrence of hidden mutations, but also to the number of steps involved in each observed mutation and to the true parental origin of the mutated allele. Besides the mode of genetic transmission, other factors influence the likelihood and magnitude of these biases, as the allele frequency distribution of the marker, the type of mutation occurred, the familial configuration considered, and even the genotypic profiles of the individuals.

The works presented in this section are based on computer simulations and delve into these factors, measuring their effect in the estimation of autosomal and X-chromosomal STR mutation rates and corresponding modelling. Corrections were suggested and an informatics tool was proposed for the correction of the incompatibilities rates obtained through the MIA for autosomal STRs. An informatics tool for X-chromosomal STRs is under construction. Computer simulations were performed using Python™ programming. Parent(s)-child duos or trios were considered for autosomal transmission, while for the X-chromosomal counterpart parent-child pedigrees based on the sex of the individuals: father-daughter, mother-daughter, and mother-son duos and father-mother-daughter trios, were independently analyzed. One million genotypic configurations were simulated for each mode of genetic transmission, pedigree, marker, population, and/or type of mutation under analysis.

- To measure and study the magnitude and type of biases obtained in mutation rates estimation relative to the allelic profile of different populations and the type of familial genotypic configuration, for autosomal genetic transmission. An informatics tool was developed (open source coded) and made freely available online at https://github.com/econdesousa/Incomp2Mut.git. This work resulted in a publication on Genes (https://doi.org/10.3390/genes13071248). It is presented in section 2.2.1. of this thesis. Due to space constraints supplementary Appendix 2 is only accessible online at https://www.mdpi.com/article/10.3390/genes13071248/s1.

- The different biases occurring in mutation rates estimation through the MIA were weighted for X-chromosomal markers in one-generational familial configurations,

for different markers and populations. These results were compared with ones obtained for autosomal markers. We are currently developing a framework to correct incompatibility rates obtained through the MIA for X-chromosomal markers, analogously to what we have developed for autosomes. This work is in final preparation as the informatics tool for correction of incompatibilities rates into mutation rates for X-chromosomal STRs is currently being developed. It is presented in section 2.2.2. of this thesis.

- To determine the frequencies of hidden mutations, and of multistep mutations that would be interpreted as single-step were computed considering X-STRs and the four X-chromosomal parent(s)-child familial configurations. One single-step and one two-step mutation (not concurrently) were randomly assigned to one of the parents' alleles in the case of trios. For trios, differential analyses were performed for either maternal or paternal meiosis. This work resulted in a publication: Antão-Sousa, S., *et al.* (2019). Underestimation and misclassification of mutations at X chromosome STRs depend on population's allelic profile. *Forensic Science International: Genetics Supplement Series*. (https://doi.org/10.1016/j.fsigss.2019.10.150) and was presented as a conference poster at the 28th Congress of the International Society for Forensic Genetics (ISFG). It is presented in section 2.2.3. of this thesis.

- To ascertain and compare the frequency with which a mutation may be ambiguously or incorrectly assigned to either the father or the mother in diploid and haplodiploid modes of genetic transmission, one single-, one two- or one three-step mutation was simulated in trios for both autosomal and X-chromosomal markers. For the latter differential analyses were performed for paternal and maternal mutations. This work resulted in a publication: Antão-Sousa, S., *et al.* (2022). How often have X- and autosomal-STRs mutations equivocal parental origin been assigned?. *Forensic Science International: Genetics Supplement Series*. (https://doi.org/10.1016/j.fsigss.2022.09.035) and was presented as an oral presentation at the 29th Congress of the International Society for Forensic Genetics (ISFG). It is presented in section 2.2.4. of this thesis.

- To compare the proportion of occurring hidden mutations or mutations interpreted as involving fewer steps for autosomal and X-chromosomal STRs, two- and three-steps mutations were simulated in every possible pedigree for both modes

of transmission, using real populations' databases. This work resulted in a publication: Antão-Sousa, S., *et al.* (2022). How frequently are Autosomal and X-STRs multistep mutations perceived as single-step?. *Forensic Science International: Genetics Supplement Series*. (https://doi.org/10.1016/j.fsigss.2022.10.022) and was presented as a conference poster at the 29th Congress of the International Society for Forensic Genetics (ISFG). It is presented in section 2.2.5. of this thesis.

2.2.1. Estimations of Mutation Rates Depend on Population Allele Frequency Distribution: The Case of Autosomal Microsatellites

Antão-Sousa, S., Conde-Sousa, E., Gusmão, L., Amorim, A., & Pinto, N. (2022). Estimations of mutation rates depend on population allele frequency distribution: the case of autosomal microsatellites. *Genes*, *13*(7), 1248. (https://doi.org/10.3390/genes13071248)

For supplementary material please check Appendix D.

genes

MDPI

Article

Estimations of Mutation Rates Depend on Population Allele Frequency Distribution: The Case of Autosomal Microsatellites

Sofia Antão-Sousa [1,2,3,4,*], Eduardo Conde-Sousa [1,5], Leonor Gusmão [4], António Amorim [1,2,3] and Nádia Pinto [1,2,6]

1 Instituto de Investigação e Inovação em Saúde (i3S), 4200-135 Porto, Portugal; econdesousa@gmail.com (E.C.-S.); aamorim@ipatimup.pt (A.A.); npinto@ipatimup.pt (N.P.)
2 Institute of Molecular Pathology and Immunology, University of Porto (IPATIMUP), 4200-465 Porto, Portugal
3 Faculty of Sciences, University of Porto (FCUP), 4169-007 Porto, Portugal
4 DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Rio de Janeiro 20550-013, Brazil; leonorbgusmao@gmail.com
5 Instituto de Engenharia Biomédica (INEB), 4200-135 Porto, Portugal
6 Center of Mathematics, University of Porto (CMUP), 4169-007 Porto, Portugal
* Correspondence: sofias@ipatimup.pt

Abstract: Microsatellites (or short-tandem repeats (STRs)) are widely used in anthropology and evolutionary studies. Their extensive polymorphism and rapid evolution make them the ideal genetic marker for dating events, such as the age of a gene or a population. This usage requires the estimation of mutation rates, which are usually estimated by counting the observed Mendelian incompatibilities in one-generation familial configurations (typically parent(s)–child duos or trios). Underestimations are inevitable when using this approach, due to the occurrence of mutational events that do not lead to incompatibilities with the parental genotypes ('hidden' or 'covert' mutations). It is known that the likelihood that one mutation event leads to a Mendelian incompatibility depends on the mode of genetic transmission considered, the type of familial configuration (duos or trios) considered, and the genotype(s) of the progenitor(s). In this work, we show how the magnitude of the underestimation of autosomal microsatellite mutation rates varies with the populations' allele frequency distribution spectrum. The Mendelian incompatibilities approach (MIA) was applied to simulated parent(s)/offspring duos and trios in different populational scenarios. The results showed that the magnitude and type of biases depend on the population allele frequency distribution, whatever the type of familial data considered, and are greater when duos, instead of trios, are used to obtain the estimates. The implications for molecular anthropology are discussed and a simple framework is presented to correct the naïf estimates, along with an informatics tool for the correction of incompatibility rates obtained through the MIA.

Keywords: microsatellites; STRs; autosomes; mutation rate estimates biases; hidden mutations; dating; evolution

Citation: Antão-Sousa, S.; Conde-Sousa, E.; Gusmão, L.; Amorim, A.; Pinto, N. Estimations of Mutation Rates Depend on Population Allele Frequency Distribution: The Case of Autosomal Microsatellites. *Genes* **2022**, *13*, 1248. https://doi.org/10.3390/genes13071248

Academic Editor: Dror Sharon

Received: 1 June 2022
Accepted: 11 July 2022
Published: 14 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Microsatellites (or short tandem repeats (STRs)) have been widely used in a wide range of scientific fields, such as population and forensic genetics, anthropology, and evolution; see, for example, [1–3]. For most of these usages, a critical parameter is required: the (germinal) mutation rate, that is, the frequency of errors in replicating DNA when producing a gamete. In the field of molecular anthropology, for example, microsatellite mutation rates are used to estimate the coalescence time between the alleles of a locus [4], the date of introduction/expansion of a variant in a population [5–8] and a variety of evolutionary events [9,10].

The main mechanism behind length mutations in microsatellites is thought to be the polymerase template slippage [11,12]. DNA strand slippage may transiently occur

during DNA synthesis, which may result in mutant products where repeat units are added or deleted within the microsatellite [11–13]. Several factors influence microsatellite mutation rates, such as the (i) allele length, (ii) repetitive motif size and sequence, and (iii) parental age and gender. Indeed, (i) longer alleles tend to have higher mutation rates [14]; (ii) longer repeats tend to have lower mutation rates and, among those with the same length, mutations vary according to their sequences [15,16]; (iii) older individuals and males present higher mutation rates [17].

Mutations that involve the gain or loss of a single repeat in the transmitted parental allele single-step mutations are assumed to be preponderant over mutations involving the gain or loss of more than one repeat (multistep mutations) [18,19]. Indeed, the most accepted mutational model is the so-called stepwise mutation model (SMM), which considers single-step mutations as the most frequent when compared to multistep mutations [20]. This bias between single and multistep mutations is supported by studies on Y microsatellites, where length mutations are undoubtedly and specifically identified in simple structure markers [21–23]. Under the SMM framework, a single-step mutation is assumed to have occurred whenever this explains the genotypic incompatibility observed in a duo or trio familial configuration. The standard method for estimating microsatellite mutation rates is detecting and quantifying Mendelian incompatibility rates in one-generation family genotypic configurations, considering either the father or the mother and the child (the so-called duos), or considering both parents and the child (trios) [22,24–27]. However, except in the case of simple structure markers in haploid systems [21,28,29], this methodology entails the underestimation of mutation rates, as mutations may not necessarily lead to incompatibilities between parent(s) and child genotypes, originating 'hidden' or 'covert' mutations [29–32] (see Figure 1 for examples).

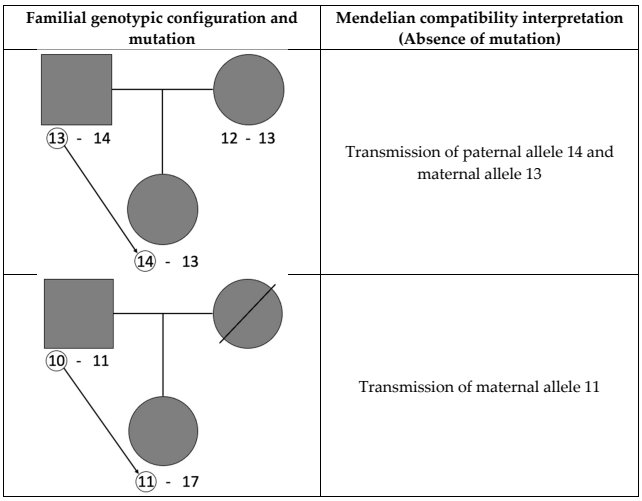


Figure 1. Examples of hidden mutations occurred in a parent–child duo and a parents–child trio, at an autosomal microsatellite. The arrows and circles indicate the allele transmission involving a mutation.

It is known that the likelihood of a mutation resulting in a Mendelian incompatibility is correlated with the type of familial configuration used [28,29], with biases being greater when duos, instead of trios, are analyzed [28–31,33]. A correction method was described previously [29], but the absence of an informatics tool for its implementation may have prevented its use.

In this work, we showed how the magnitude of the underestimation of autosomal microsatellite mutation rates varies with the populations' allele frequency distribution spectrum, using simulated parent(s)–child duos and trios in different populational scenarios

and assuming single-step mutations. Simulated familial duos and trios were generated considering a single-step mutation for each familial clustering and marker, using real and theoretical mock allele frequency distributions (henceforth called mock). Mock allele frequency distributions were considered to diversify the analyzed population allelic backgrounds and were designed by us considering pre-defined distributions. The populations/markers showing the highest rates of hidden mutations were assumed to be those with the greatest mutation rate biases when a standard approach to quantifying the Mendelian incompatibilities (Mendelian incompatibilities approach (MIA)) is used.

We aim to study the magnitude and type of biases obtained in mutation rate estimates, depending on the population allele frequency distribution and the type of familial data that are considered.

The implications for molecular anthropology are highlighted and a simple framework to correct the naïf mutation estimates, obtained through an analysis of Mendelian incompatibilities, is presented, along with a user-friendly and freely available informatics tool for the corresponding correction of incompatibility rates to mutation rates.

2. Materials and Methods

Familial genotypic configurations, mother–child or father–child duos, and mother–father–child trios, were generated by resorting to Python™ programming language. Since we aimed to measure the proportion of hidden mutations present in different population backgrounds, parental genotypes were randomly attributed from both real and mock population allele frequency distributions. Real allelic distributions concern ten autosomal microsatellites: CSF1PO, D1S1656, D21S11, D2S441, D3S1358, FGA, SE33, TH01, TPOX, and VWA, for the Norway, Somalia, and Spain populations [34]—see Supplementary Material File S1 for a graphic representation of the allelic distributions and Table S1 for population information (size, expected heterozygosity, polymorphism informative content (PIC) allele number, and allelic range). These markers were selected due to their distinct allelic distributions. On the other hand, to diversify the scenarios obtained from real populations and forensic markers, six mock (predefined) frequency distributions were designed by us: normal, bimodal, and constant distributions. For each case, narrow and wide allelic spans (with 10 and 20 alleles, respectively) were considered (Figure 2).

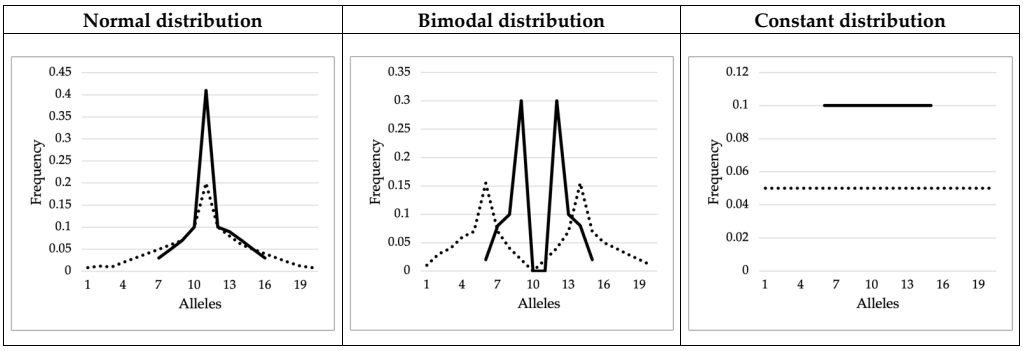


Figure 2. Mock allele frequencies, considering predefined distributions. The full lines correspond to the situation where 10 alleles are considered and the dotted lines to the cases with 20 alleles (narrow and wide distributions, respectively).

Parental genotypic configurations were generated by considering the allele frequencies of Norway, Somalia, and Spain populations [34] and the allele frequencies of the mock distributions (Figure 2). For each marker and population database, 1,000,000 familial genotypic configurations (duos and trios) were simulated, assuming, for each case, the occurrence of exactly one single-step mutation. To simulate the parental alleles, a cumulative relative frequency associated with each allele was considered for each marker. The two and four

parental alleles considered in the case of duos and trios, respectively, were obtained considering two and four, resp., random numbers between 0 and 1. The corresponding allele was then chosen considering this number; the greater the frequency of the allele, the most likely that the allele was selected. When trios were simulated, the meiosis suffering mutation (either paternal or maternal) was randomly selected. Mutated alleles were assumed to be transmitted to the offspring, while the other filial allele was randomly selected either from the population (in the case of duos) or the other parent (in the case of trios).

As a parental mutated allele was assumed to be transmitted to the offspring in all the cases, there were two possible outcomes for each simulated familial configuration: the familial genotypic configuration was incompatible with the Mendelian inheritance, or otherwise. Under the standard, general, approach, the rate of the cases resulting in the first for a specific marker would be presented as the marker specific average mutation rate; while the rate of cases that resulted in the latter equates to the rate of hidden mutations, which would remain unnoticed.

The rate of hidden mutations was quantified for the different markers, populations, and familial configurations, assuming that the higher the rate, the greater the bias of the corresponding marker-specific mutation rate estimated through Mendelian incompatibilities.

Linear regression analyses were performed using Microsoft Excel®. The heterozygosity of each marker was calculated as $1 - \sum_i p_i^2$, where p_i is the frequency of the allele i.

Fisher Exact tests to ascertain p values were computed considering a level of significance equal to 0.05. Python algorithms are openly available at https://github.com/econdesousa/Incomp2Mut.git (accessed on 23 March 2022), along with an informatics tool.

To replicate the simulations described in this work, and obtain a corrected mutation rate estimate for any marker, population, or familial configuration, the algorithms in https://github.com/econdesousa/Incomp2Mut.git (accessed on 23 March 2022) must be run for the target marker and incompatibility rates obtained through MIA. An example file on how to present the allelic distributions and a detailed explanation in video format of how to proceed with the informatic analyses are provided.

3. Results

In this section, results are presented comparing the accuracy of the estimates obtained for autosomal STR mutation rates, considering the analysis of both familial duos and trios, for different markers, populations, and mock allele frequencies, through the evaluation of the hidden mutation rates that were observed—see Table 1 and Figure 3.

3.1. Real Allele Frequency Distributions

As expected, parent–child duos concealed single-step mutations more often than parent–child trios (Table 1). These biases were strongly dependent on the allele frequency distribution of the marker. As a striking example, when considering the population of Norway, the rate of hidden mutations obtained for trios in marker CSF1PO were greater than the one obtained for marker D1S1656 when considering duos. This indicates that, for that population and markers, better estimates are expected for the D1S1656 mutation rate when duos are studied than for CSF1PO when trios are used once the same number of meiosis are analyzed. Indeed, within each of the three considered population databases [34], widely different proportions of hidden mutations were found for the 10 analyzed markers, even when the same familial configurations were used (either duos or trios). For example, in parent–child duos, considering the Norwegian database, mutations were concealed 4.3 more often for TPOX than for SE33. Globally, the proportion of hidden mutations varied from 5.4% (in SE33, for the Norwegian population, using trios) to 62.2% (in TPOX, also for the Norwegian population, when using duos). The ratio of hidden mutations found for the 10 analyzed markers in each population was computed. Statistically significant differences were found for virtually all pairwise comparisons; see Supplementary Material File S2. Within all populations, the standard deviation in the ratios of marker-specific hidden mutations was shown to be high, varying between 0.058 (for Somalia, in trios), and 0.142

(for Norway, in duos); see Table S2. Our results show that distinct levels of confidence for mutation rate estimates are expected for different markers within the same population; thus, marker-specific mutation rates should be estimated.

Furthermore, marker-specific pairwise analyses were computed, comparing the proportion of hidden mutations that were obtained for the different populations we studied; see Supplementary Material File S2. Most (98.6%, $\alpha = 0.000115$) pairwise population comparisons showed that the ratio of hidden mutations obtained for a specific marker significantly differed from population to population.

Therefore, we conclude that the difference between incompatibility rates (obtained through observations of Mendelian incompatibilities in duos or trios) and mutation rates depends on the allele frequency distribution in the population, whatever the familial configuration used (duos or trios). Globally, the markers that showed the worst and best mutation rate estimates (highest and smallest rates of hidden mutations, respectively) were D3S1358 and SE33, respectively.

None of the three populations we analyzed showed a consistent lowest value of hidden mutations across all markers, with most markers showing statistically significant pairwise differences for the rate of hidden mutations when analyzed in different populations; see Supplementary Material File S2. Nevertheless, the standard deviation associated with the marker-specific hidden mutation rates analyzed in different populations was small (maximum average $\sigma = 0.031$, see Table S3), in contrast with the one found when different markers were analyzed within a specific population (maximum average $\sigma = 0.124$; see Table S2).

Table 1. Rates of hidden mutations per marker (real and mock allelic distributions) and familial configuration considered (either duos or trios). One single-step mutation was simulated in one, randomly selected, parental meiosis of each of the 1,000,000 parent–child duos and parents–child trios, considering the allelic distributions of 10 autosomal STRs for the populations of Norway, Somalia, and Spain [34] and the allelic distributions of the 6 artificially generated markers. In the latter, N refers to the number of alleles in the marker.

Markers			Duos			Trios		
			Norway	Somalia	Spain	Norway	Somalia	Spain
Allele Frequencies	Real allelic distributions	**CSF1PO**	**0.546**	**0.497**	**0.497**	**0.269**	**0.219**	**0.207**
		D1S1656	0.249	0.312	0.325	0.087	0.128	0.135
		D21S11	0.372	0.348	0.35	0.163	0.15	0.147
		D2S441	0.497	0.419	0.393	0.141	0.125	0.108
		D3S1358	0.463	0.526	0.542	0.224	0.259	0.271
		FGA	0.334	0.329	0.328	0.15	0.145	0.146
		SE33	0.143	0.178	0.171	0.054	0.069	0.065
		TH01	0.452	0.508	0.521	0.102	0.227	0.233
		TPOX	0.622	0.508	0.525	0.119	0.192	0.198
		VWA	0.445	0.425	0.435	0.213	0.205	0.205
	Mock allelic distributions	**Normal (N = 10)**	0.696			0.157		
		Normal (N = 20)	0.614			0.09		
		Bimodal (N = 10)	0.663			0.132		
		Bimodal (N = 20)	0.586			0.066		
		Constant (N = 10)	0.63			0.105		
		Constant (N = 20)	0.57			0.052		

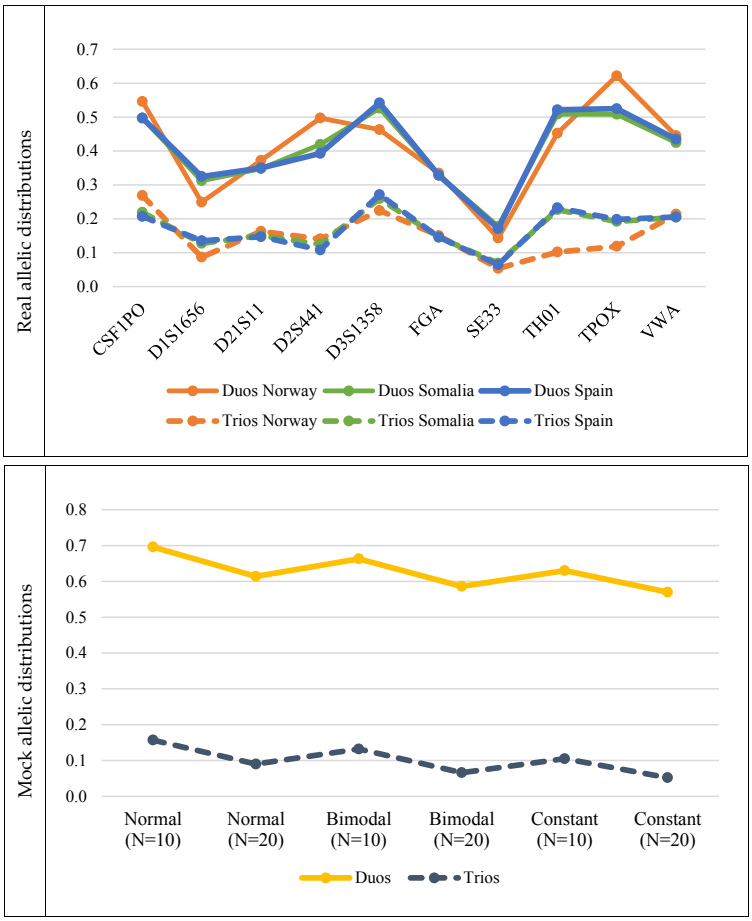


Figure 3. Graphical representations of the proportion of hidden mutations per marker (upper for real population distributions and lower for mock ones, considering the indicated distribution) and familial configuration. Full lines connect the dots corresponding to the proportion of hidden mutations for each marker in duos; dotted lines connect the dots corresponding to the proportion of hidden mutations for each marker in trios. For the mock distributions, N refers to the number of alleles considered in the marker. For example, "Normal (N = 10)" refers to the mock marker designed with a normal and narrow distribution, with 10 alleles, whereas "Normal (N = 20)" refers to the mock marker designed with a normal and wider distribution, with 20 alleles.

3.2. Mock Allele Frequency Distributions

The six mock-allelic distributions: normal (narrow and wide), bimodal (narrow and wide), and constant (narrow and wide), showed widely different proportions of hidden mutations. More specifically, in parent–child duos simulated considering the normal distribution and 10 alleles, mutations were concealed 4.4 more times than in trios. Globally, the proportion of hidden mutations varied from 52.0% (in the wide constant distribution, using trios) to 69.6% (in the narrow normal distribution, using duos). As before, our results show that distinct levels of confidence for mutation rates estimates are expected for markers with different allelic distributions, again with duos hiding more mutations than trios. Besides, markers with a narrower distribution, i.e., with fewer alleles, hid more mutations than markers with more alleles. The ratio of hidden mutations was greater when the number of alleles increased for all the analyzed distributions. For example, for a normal, unimodal distribution with 10 markers (narrow distribution), duos concealed 4.4 times more muta-

tions than trios. This figure increased to 6.8 times when 20 alleles were considered. This shows that biases resulting from the analysis of duos for estimating mutation rates via the computation of Mendelian incompatibility rates may be greater in less polymorphic populations. There is also a linear correlation between the expected heterozygosity and the rate of hidden mutations observed for the six mock allelic distributions for both duos and trios ($r^2 = 0.9841$ and $r^2 = 0.9912$, respectively); see Figure 4. Nevertheless, for the real allelic distributions that were studied, this high correlation was only verified for the case of duos ($r^2 = 0.911$; and $r^2 = 0.4609$ for trios).

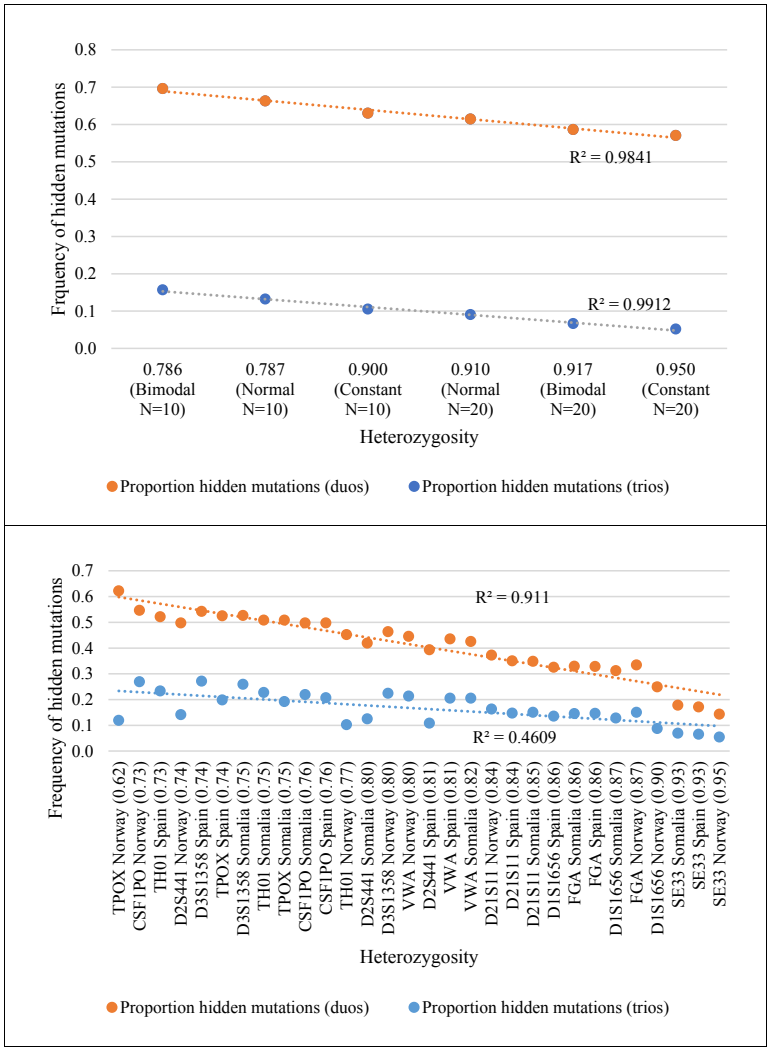


Figure 4. Graphical representations of the correlation between the frequency of hidden mutations observed per marker (upper for real population distributions and lower for mock ones, considering the indicated distribution) and markers' heterozygosity. Orange corresponds to the correlation of hidden mutations and the heterozygosity of each marker in duos, and blue in trios. For the mock distributions, N refers to the number of alleles considered in the marker. For example, "Normal (N = 10)" refers to the mock marker designed with a normal and narrow distribution, with 10 alleles, whereas "Normal (N = 20)" refers to the mock marker designed with a normal and wider distribution, with 20 alleles. Heterozygosity was calculated as $1 - \sum_i p_i^2$, where p_i is the frequency of the allele i.

4. Discussion

The accuracy of autosomal mutation rate estimates obtained through Mendelian incompatibilities varies between markers and populations according to allele frequency distributions, whatever the type of one-generation family data (parent(s)–child duos or trios) employed. Since mutations do not necessarily lead to Mendelian incompatibilities, this approach inherently underestimates their frequency.

It was previously acknowledged that the mutation rate estimates obtained through the observation of Mendelian incompatibilities at autosomal and X-chromosomal transmissions imply biases [28–31,33], and some procedures to correct them were already published for autosomal transmission [29–31,33]. Despite this, the generally accepted approach continues to be the direct estimation of mutation rates through the counting of Mendelian incompatibilities, without any correction; see, for example [24–27,35–38].

In autosomes, biases are more important in parent–child duos than in trios [29], showing that pooling data from the two types of sources is not acceptable, as it prevents any kind of a posteriori correction.

We have demonstrated that the probability of the occurrence of hidden mutations depends on the allele frequency distribution; therefore, the same marker may show different estimates in distinct populations despite the mutation rate value being the same. At this point, it should be highlighted that the real frequency distributions available for our analyses correspond to the markers designed for forensic individual identification, comprising STRs with a high diversity within, rather than between, populations. This is not the case for markers of anthropological interest, which were selected to maximize the differences between populations. In this case, pooling data from different populations to estimate mutation rates should be carefully thought out and planned, as different allelic distributions carry different likelihoods of disclosing mutations.

The framework we present in this work, described in the Materials and Methods section and thoroughly explained in https://github.com/econdesousa/Incomp2Mut.git (accessed on 23 March 2022), can be used to correct the mutation rates estimated through the MIA. To obtain the proposed corrective factor, it is only necessary to know the distribution of the allele frequencies at the loci of interest, the incompatibility rate observed and whether duos or trios were used to ascertain said rate. The output will be the corrected (for hidden mutations) mutation rate for the analyzed microsatellite. As exemplified, if parent–child duos are used and an R rate of Mendelian incompatibilities is found at the CSF1PO locus in the Norwegian population, the corrected value of R/(1–0.546) should be used as the estimated mutation rate at this locus and population, which represents nearly double the value estimated via MIA.

5. Conclusions

The accuracy of microsatellite mutation rate estimates obtained through the observation of Mendelian incompatibilities in parent(s)–child duos or trios depends on several factors, including the population allele frequency distribution. We showed that even when trios are used, as many as 27.1% (as obtained for marker D3S1358 for the population of Spain) of the mutations did not lead to any incompatibility. Although we framed our analyses under the knowledge that single-step mutations are the most frequent, the magnitude and the types of the biases increase if other mutations are considered (data not shown). We also did not consider the causes of the evidenced differences in the estimates across populations, i.e., whether biases are due to simple statistical properties of allelic distributions or intrinsic differences in allelic mutability. Whatever the reasons for these differences, its effect is the same on the estimation accuracy, as it only depends on the allele frequency distribution. However, the long-term evolutionary consequences are different. It is also important to note that population differentiation is expected to be lower for autosomal than for heterosomal markers, which are more susceptible to genetic drift. Therefore, the variation observed in the frequency of hidden mutations among populations when using autosomal markers may be even higher for heterosomal microsatellites.

The impact of this systematic underestimation inherent to the approach can be particularly burdensome in crucial anthropological problems, such as when dating evolutionary events; see, for example, [6,7,10].

We propose a simple method to obtain mutation rate estimates from Mendelian incompatibilities when using familial duos or trios, aiming to minimize the impact of hidden mutations. An informatics tool is provided at https://github.com/econdesousa/Incomp2Mut.git (accessed on 23 March 2022) to replicate this approach and obtain corrected mutation rates for any autosomal microsatellite, employing the allele frequency distribution and incompatibility rate estimated for either duos or trios.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes13071248/s1, Supplementary Material File S1: Allelic distribution of marker CSF1PO for the populations of Norway (N = 19,156), Somalia (N = 1598) and Spain (N = 2500); Supplementary Material File S2: *p* values found in pairwise comparisons between the frequency of hidden mutations found when one-step mutation is simulated in parents-child trios for each marker and population. Table S1: Population size (N), expected heterozygosity (with Nei correction), polymorphic informative content (PIC), number of alleles and allelic range per marker and population; Table S2: Populations showing the highest and lowest standard deviations (σ between parentheses) considering the proportion of hidden mutations across markers, for each familial configuration type. A single-step mutation was simulated in one of the parental meiosis of 1,000,000 configurations of each type, considering the allele frequencies of 10 autosomal STRs in three populations (Norway, Somalia, and Spain); Table S3: Average standard deviation and markers showing the highest and lowest values (in parentheses) of the proportion of hidden mutations across the different population databases: Norway, Somalia, and Spain, for each familial configuration type. A single-step mutation was simulated in one of the parental meiosis of 1,000,000 configurations of each type, considering the allele frequencies of 10 autosomal STRs in three populations (Norway, Somalia, and Spain).

Author Contributions: Conceptualization, L.G., A.A. and N.P.; Data curation, S.A.-S., E.C.-S. and N.P.; Formal analysis, S.A.-S., E.C.-S., L.G., A.A. and N.P.; Funding acquisition, S.A.-S., E.C.-S., L.G., A.A. and N.P.; Investigation, S.A.-S., E.C.-S., L.G., A.A. and N.P.; Methodology, S.A.-S., E.C.-S., L.G., A.A. and N.P.; Software, S.A.-S., E.C.-S. and N.P.; Supervision, L.G., A.A. and N.P.; Validation, S.A.-S., E.C.-S., L.G., A.A. and N.P.; Writing–original draft, S.A.-S.; Writing–review and editing, S.A.-S., E.C.-S., L.G., A.A. and N.P. All authors have read and agreed to the published version of the manuscript.

Funding: FEDER—Fundo Europeu de Desenvolvimento Regional funds through the COMPETE 2020—Operacional Programme for Competitiveness and Internationalisation (POCI), Portugal 2020; FCT—Fundação para a Ciência e Tecnologia: projects "Institute for Research and Innovation in Health Sciences" (POCI-01-0145-FEDER-007274); FCT—Fundação para a Ciência e Tecnologia: Decree-Law no.57/2016 of August 29; FCT—Fundação para a Ciência e Tecnologia: SFRH/BD/136284/2018; CNPq—National Council for Scientific and Technological Development: ref. 306342/2019-7; FAPERJ—Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro: (CNE-2018); i3S—Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal: POCI-01-0145-FEDER-007274; CMUP—Center of Mathematics of the University of Porto: UIDB/00144/2020; PPBI—Portuguese Platform of BioImaging: PPBI-POCI-01-0145-FEDER-022122.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Flores-Espinoza, R.; Paz-Cruz, E.; Ruiz-Pozo, V.A.; Lopez-Carrera, M.; Cabrera-Andrade, A.; Gusmão, L.; Burgos, G. Investigating genetic diversity in admixed populations from Ecuador. *Am. J. Phys. Anthropol.* **2021**, *176*, 109–119. [CrossRef]
2. Srithawong, S.; Muisuk, K.; Srikummool, M.; Kampuansai, J.; Pittayaporn, P.; Ruangchai, S.; Liu, D.; Kutanan, W. Close genetic relationship between central Thai and Mon people in Thailand revealed by autosomal microsatellites. *Int. J. Leg. Med.* **2021**, *135*, 445–448. [CrossRef] [PubMed]
3. He, G.; Wang, Z.; Wang, M.; Hou, Y. Genetic Diversity and Phylogenetic Differentiation of Southwestern Chinese Han: A comprehensive and comparative analysis on 21 non-CODIS STRs. *Sci. Rep.* **2017**, *7*, 13730. [CrossRef] [PubMed]
4. Sun, J.X.; Mullikin, J.C.; Patterson, N.; Reich, D.E. Microsatellites are molecular clocks that support accurate inferences about history. *Mol. Biol. Evol.* **2009**, *26*, 1017–1027. [CrossRef] [PubMed]
5. Peixoto, A.; Santos, C.; Pinheiro, M.; Pinto, P.; Soares, M.J.; Rocha, P.; Gusmão, L.; Amorim, A.; Van Der Hout, A.; Gerdes, A.-M.; et al. International distribution and age estimation of the Portuguese BRCA2 c.156_157insAlu founder mutation. *Breast Cancer Res. Treat.* **2011**, *127*, 671–679. [CrossRef]
6. Sharony, R.; Martins, S.; Costa, I.P.D.; Zaltzman, R.; Amorim, A.; Sequeiros, J.; Gordon, C.R. Yemenite-Jewish families with Machado–Joseph disease (MJD/SCA3) share a recent common ancestor. *Eur. J. Hum. Genet.* **2019**, *27*, 1731. [CrossRef]
7. Martins, S.; Soong, B.-W.; Wong, V.C.N.; Giunti, P.; Stevanin, G.; Ranum, L.P.W.; Sasaki, H.; Riess, O.; Tsuji, S.; Coutinho, P.; et al. Mutational origin of Machado-Joseph disease in the Australian Aboriginal communities of Groote Eylandt and Yirrkala. *Arch. Neurol.* **2012**, *69*, 746–751. [CrossRef]
8. Martins, S.; Calafell, F.; Gaspar, C.; Wong, V.C.N.; Silveira, I.; Nicholson, G.A.; Brunt, E.R.; Tranebjaerg, L.; Stevanin, G.; Hsieh, M.; et al. Asian origin for the worldwide-spread mutational event in Machado-Joseph disease. *Arch. Neurol.* **2007**, *64*, 1502–1509. [CrossRef]
9. Boattini, A.; Sarno, S.; Mazzarisi, A.M.; Viroli, C.; De Fanti, S.; Bini, C.; Larmuseau, M.H.D.; Pelotti, S.; Luiselli, D. Estimating Y-Str Mutation Rates and Tmrca Through Deep-Rooting Italian Pedigrees. *Sci. Rep.* **2019**, *9*, 9032. [CrossRef]
10. Regueiro, M.; Alvarez, J.; Rowold, D.; Herrera, R.J. On the origins, rapid expansion and genetic diversity of Native Americans from hunting-gatherers to agriculturalists. *Am. J. Phys. Anthr.* **2013**, *150*, 333–348. [CrossRef]
11. Strand, M.; Prolla, T.A.; Liskay, R.M.; Petes, T.D. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* **1993**, *365*, 274–276. [CrossRef] [PubMed]
12. Schlötterer, C.; Tautz, D. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.* **1992**, *20*, 211. [CrossRef] [PubMed]
13. Eckert, K.A.; Hile, S.E. Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. *Mol. Carcinog.* **2009**, *48*, 379–388. [CrossRef]
14. Wierdl, M.; Dominska, M.; Petes, T.D. Microsatellite instability in yeast: Dependence on the length of the microsatellite. *Genetics* **1997**, *146*, 769–779. [CrossRef] [PubMed]
15. Brinkmann, B.; Klintschar, M.; Neuhuber, F.; Hühne, J.; Rolf, B. Mutation Rate in Human Microsatellites: Influence of the Structure and Length of the Tandem Repeat. *Am. J. Hum. Genet.* **1998**, *62*, 1408–1415. [CrossRef]
16. Ballantyne, K.N.; Goedbloed, M.; Fang, R.; Schaap, O.; Lao, O.; Wollstein, A.; Choi, Y.; van Duijn, K.; Vermeulen, M.; Brauer, S.; et al. Mutability of Y-Chromosomal Microsatellites: Rates, Characteristics, Molecular Bases, and Forensic Implications. *Am. J. Hum. Genet.* **2010**, *87*, 341. [CrossRef]
17. Sun, J.X.; Helgason, A.; Masson, G.; Ebenesersdóttir, S.S.; Li, H.; Mallick, S.; Gnerre, S.; Patterson, N.; Kong, A.; Reich, D.; et al. A direct characterization of human mutation based on microsatellites. *Nat. Genet.* **2012**, *44*, 1161–1165. [CrossRef]
18. Xu, X.; Peng, M.; Fang, Z.; Xu, X. The direction of microsatellite mutations is dependent upon allele length. *Nat. Genet.* **2000**, *24*, 396–399. [CrossRef]
19. Weber, J.L.; Wong, C. Mutation of human short tandem repeats. *Hum. Mol. Genet.* **1993**, *2*, 1123–1128. [CrossRef]
20. Kimura, M.; Ohta, T. Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc. Natl. Acad. Sci. USA* **1978**, *75*, 2868. [CrossRef]
21. Pinto, N.; Gusmão, L.; Amorim, A. Mutation and mutation rates at y chromosome specific Short Tandem Repeat Polymorphisms (STRs): A reappraisal. *Forensic Sci. Int. Genet.* **2014**, *9*, 20–24. [CrossRef] [PubMed]
22. Antão-Sousa, S.; Sánchez-Diz, P.; Abovich, M.; Alvarez, J.; Carvalho, E.; Silva, C.; Domingues, P.; Farfán, M.; Gutierrez, A.; Pontes, L.; et al. Mutation rates and segregation data on 16 Y-STRs: An update to previous GHEP-ISFG studies. *Forensic Sci. Int. Genet. Suppl. Ser.* **2017**, *6*, e601–e602. [CrossRef]
23. Dupuy, B.M.; Stenersen, M.; Egeland, T.; Olaisen, B. Y-chromosomal microsatellite mutation rates: Differences in mutation rate between and within loci. *Hum. Mutat.* **2004**, *23*, 117–124. [CrossRef] [PubMed]
24. Jin, B.; Su, Q.; Luo, H.; Li, Y.; Wu, J.; Yan, J.; Hou, Y.; Liang, W.; Zhang, L. Mutational analysis of 33 autosomal short tandem repeat (STR) loci in southwest Chinese Han population based on trio parentage testing. *Forensic Sci. Int. Genet.* **2016**, *23*, 86–90. [CrossRef] [PubMed]
25. Pinto, N.; Pereira, V.; Mas, C.T.; Loiola, S.; Carvalho, E.F.; Modesti, N.; Maxzud, M.; Marcucci, V.; Cano, H.; Cicarelli, R.; et al. Paternal and maternal mutations in X-STRs: A GHEP-ISFG collaborative study. *Forensic Sci. Int. Genet.* **2020**, *46*, 102258. [CrossRef]
26. García, M.G.; Catanesi, C.I.; Penacino, G.A.; Gusmão, L.; Pinto, N. X-chromosome data for 12 STRs: Towards an Argentinian database of forensic haplotype frequencies. *Forensic Sci. Int. Genet.* **2019**, *41*, e8–e13. [CrossRef]

27. Sun, H.; Liu, S.; Zhang, Y.; Whittle, M.R. Comparison of southern Chinese Han and Brazilian Caucasian mutation rates at autosomal short tandem repeat loci used in human forensic genetics. *Int. J. Leg. Med.* **2014**, *128*, 1–9. [CrossRef]
28. Antão-Sousa, S.; Conde-Sousa, E.; Gusmão, L.; Amorim, A.; Pinto, N. Underestimation and misclassification of mutations at X chromosome STRs depend on population's allelic profile. *Forensic Science International: Genet. Suppl. Ser.* **2019**, *7*, 718–720. [CrossRef]
29. Slooten, K.; Ricciardi, F. Estimation of mutation probabilities for autosomal STR markers. *Forensic Sci. Int. Genet.* **2013**, *7*, 337–344. [CrossRef]
30. Chakraborty, R.; Stivers, D.N.; Zhong, Y. Estimation of mutation rates from parentage exclusion data: Applications to STR and VNTR loci. *Mutat. Res.* **1996**, *354*, 41–48. [CrossRef]
31. Vicard, P.; Dawid, A.P. A statistical treatment of biases affecting the estimation of mutation rates. *Mutat. Res.* **2004**, *547*, 19–33. [CrossRef] [PubMed]
32. Brenner, C.H. Multiple mutations, covert mutations and false exclusions in paternity casework. *Int. Congr. Ser.* **2004**, *1261*, 112–114. [CrossRef]
33. Vicard, P.; Dawid, A.P.; Mortera, J.; Lauritzen, S.L. Estimating mutation rates from paternity casework. *Forensic Sci. Int. Genet.* **2008**, *2*, 9–18. [CrossRef]
34. Kling, D.; Tillmar, A.O.; Egeland, T. Familias 3—Extensions and new functionality. *Forensic Sci. Int. Genet.* **2014**, *13*, 121–127. [CrossRef] [PubMed]
35. Lan, Q.; Wang, H.; Shen, C.; Guo, Y.; Yin, C.; Xie, T.; Fang, Y.; Zhou, Y.; Zhu, B. Mutability analysis towards 21 STR loci included in the AGCU 21 + 1 kit in Chinese Han population. *Int. J. Leg. Med.* **2018**, *132*, 1287–1291. [CrossRef] [PubMed]
36. Liu, Q.L.; Chen, Y.F.; Huang, X.L.; Liu, K.Y.; Zhao, H.; Lu, D.J. Population data and mutation rates of 19 STR loci in seven provinces from China based on GoldeneyeTM DNA ID System 20A. *Int. J. Leg. Med.* **2017**, *131*, 653–656. [CrossRef] [PubMed]
37. Hongdan, W.; Bing, K.; Ning, S.; Miao, H.; Bo, Z.; Yuxin, G.; Bofeng, Z.; Shixiu, L.; Zhaoshu, Z. Evaluation of the genetic parameters and mutation analysis of 22 STR loci in the central Chinese Han population. *Int. J. Leg. Med.* **2017**, *131*, 103–105. [CrossRef]
38. Zhao, Z.; Zhang, J.; Wang, H.; Liu, Z.-P.; Liu, M.; Zhang, Y.; Sun, L.; Zhang, H. Mutation rate estimation for 15 autosomal STR loci in a large population from Mainland China. *Meta Gene* **2015**, *5*, 150–156. [CrossRef]

Appendix D

Antão-Sousa, S., Conde-Sousa, E., Gusmão, L., Amorim, A., & Pinto, N. (2022). Estimations of mutation rates depend on population allele frequency distribution: the case of autosomal microsatellites. *Genes*, *13*(7), 1248. (https://doi.org/10.3390/genes13071248)

File S1

Figure 1: Allelic distribution of marker CSF1PO for the populations of Norway (N = 19156), Somalia (N = 1598) and Spain (N = 2500).

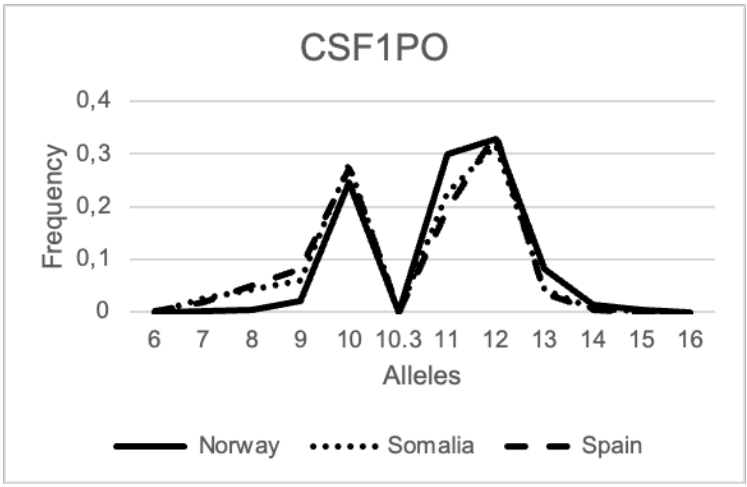


Figure 2: Allelic distribution of marker D1S1656 for the populations of Norway (N = 3472), Somalia (N = 488) and Spain (N = 2500).

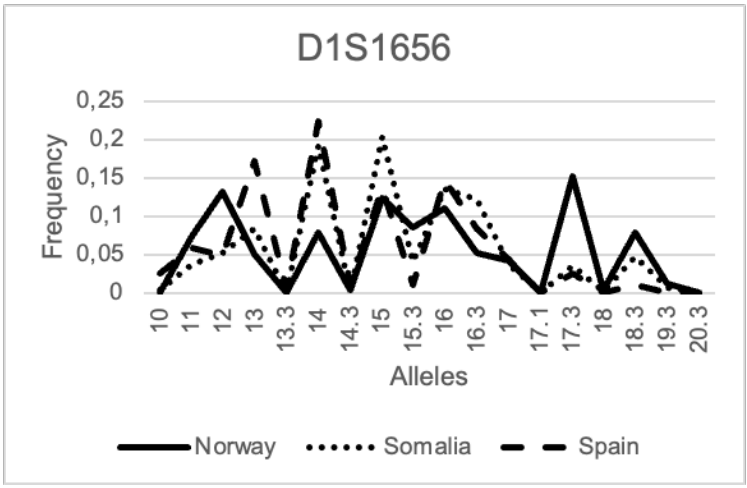


Figure 3: Allelic distribution of marker D2S441 for the populations of Norway (N = 3472), Somalia (N = 488) and Spain (N = 2500).

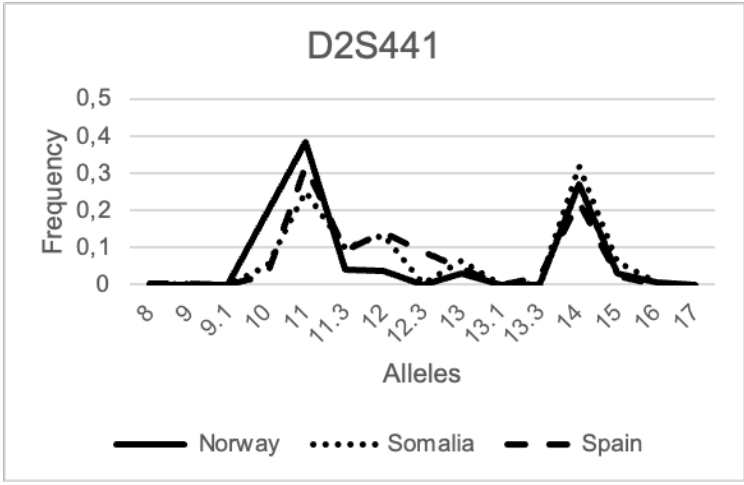


Figure 4: Allelic distribution of marker D3S1358 for the populations of Norway (N = 19172), Somalia (N = 1598) and Spain (N = 2500).

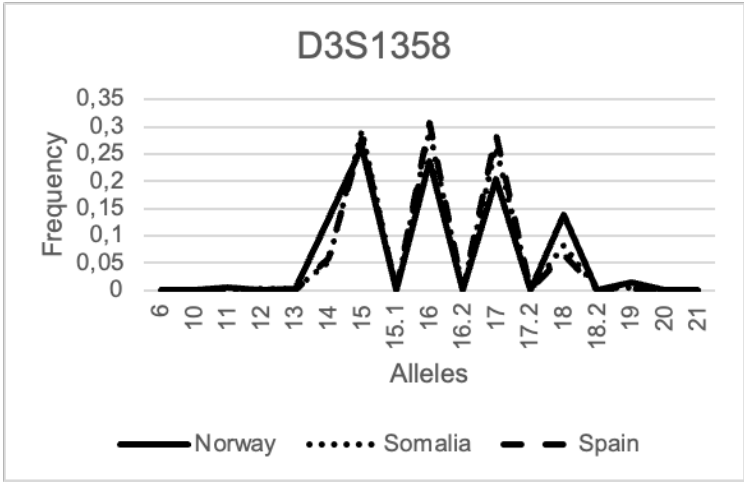


Figure 5: Allelic distribution of marker D21S11 for the populations of Norway (N = 19170), Somalia (N = 1598) and Spain (N = 2500).

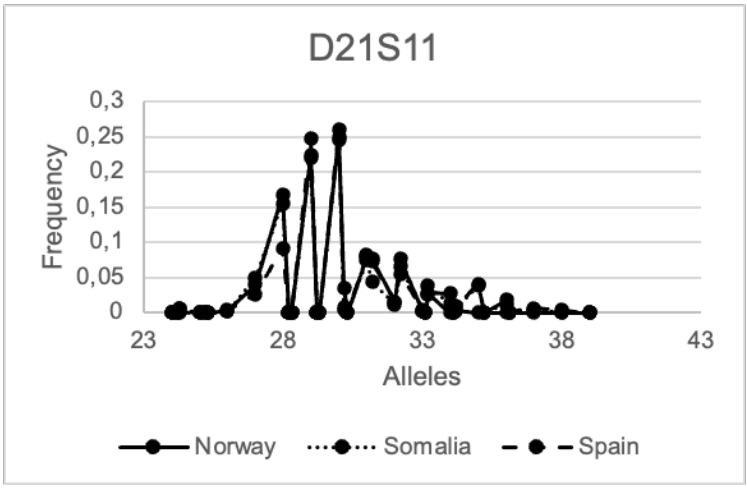


Figure 6: Allelic distribution of marker FGA for the populations of Norway (N = 19164), Somalia (N = 1598) and Spain (N = 2500).

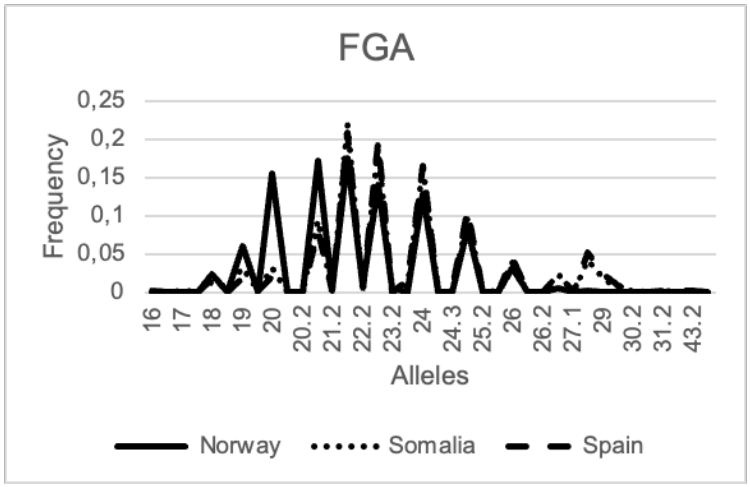


Figure 7: Allelic distribution of marker SE33 for the populations of Norway (N = 6318), Somalia (N = 1348) and Spain (N = 2500).

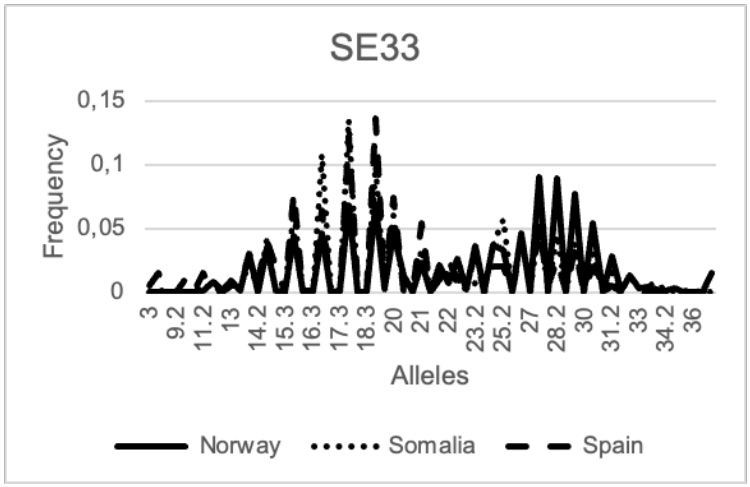


Figure 8: Allelic distribution of marker TH01 for the populations of Norway (N = 19172), Somalia (N = 1598) and Spain (N = 2500).

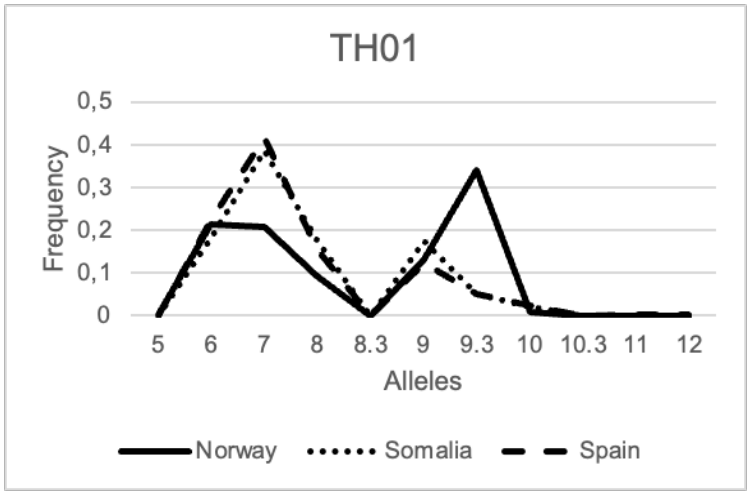


Figure 9: Allelic distribution of marker TPOX for the populations of Norway (N = 19162), Somalia (N = 1596) and Spain (N = 2500).

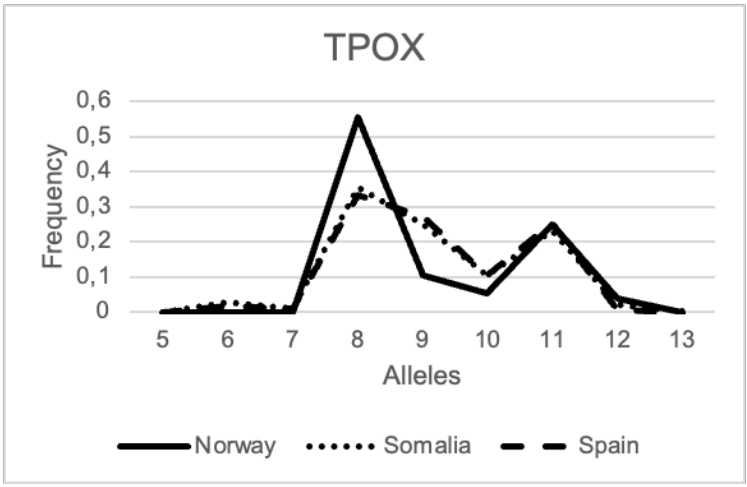
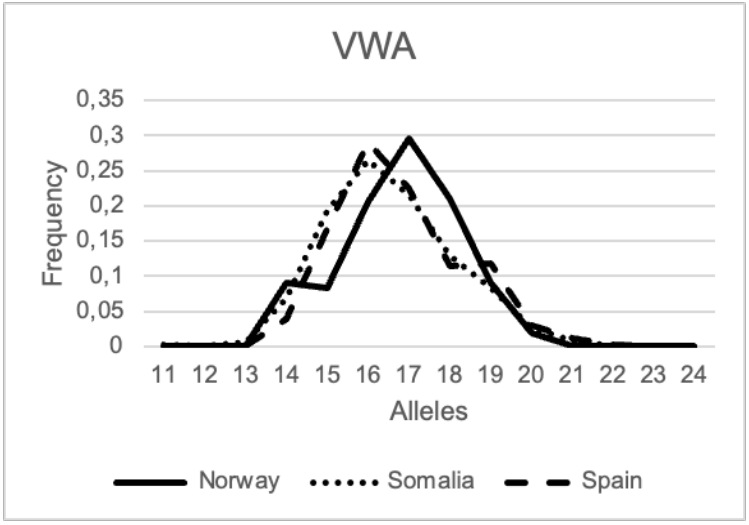


Figure 10: Allelic distribution of marker VWA for the populations of Norway (N = 19170), Somalia (N = 1597) and Spain (N = 2500).



Appendix 2: *p* values found in pairwise comparisons between the frequency of hidden mutations found when one-step mutation is simulated in parents-child trios for each marker and population. It can be assessed at https://www.mdpi.com/article/10.3390/genes13071248/s1.

Table S1: Population size (N), expected heterozygosity (with Nei correction), polymorphic informative content (PIC), number of alleles and allelic range per marker and population.

Marker and population	N	Heterozygosity (Nei correction)	Polymorphic Informative Content (PIC)	Number of alleles	Allelic range
CSF1PO (Norway)	19156	0.7335	0.6860	8	9
CSF1PO (Somalia)	1598	0.7626	0.7250	7	8
CSF1PO (Spain)	2500	0.7609	0.7241	8	8
D1S1656 (Norway)	3472	0.8992	0.8905	9	10.3
D1S1656 (Somalia)	488	0.8699	0.8568	8	9.3
D1S1656 (Spain)	2500	0.8636	0.8493	8	8.3
D21S11 (Norway)	19170	0.8377	0.8186	10	11.2
D21S11 (Somalia)	1598	0.8489	0.8327	9	14.7
D21S11 (Spain)	2500	0.8432	0.8271	9	13.7
D2S441 (Norway)	3472	0.7361	0.6951	11	8
D2S441 (Somalia)	488	0.7955	0.7687	10	7
D2S441 (Spain)	2500	0.8072	0.7835	11	9
D3S1358 (Norway)	19172	0.7957	0.7644	12	11
D3S1358 (Somalia)	1598	0.7496	0.7065	12	9
D3S1358 (Spain)	2500	0.7397	0.6934	12	13

Marker and population	N	Heterozygosity (Nei correction)	Polymorphic Informative Content (PIC)	Number of alleles	Allelic range
FGA (Norway)	19164	0.8666	0.8520	13	12.8
FGA (Somalia)	1598	0.8640	0.8501	14	28.2
FGA (Spain)	2500	0.8642	0.8501	12	27.2
SE33 (Norway)	6318	0.9479	0.9454	17	37.8
SE33 (Somalia)	1348	0.9336	0.9299	14	29.7
SE33 (Spain)	2500	0.9349	0.9313	14	28.2
TH01 (Norway)	19172	0.7683	0.7325	24	6
TH01 (Somalia)	1598	0.7501	0.7134	23	6
TH01 (Spain)	2500	0.7338	0.6962	22	6
TPOX (Norway)	19162	0.6157	0.5662	31	8
TPOX (Somalia)	1596	0.7458	0.7040	26	7
TPOX (Spain)	2500	0.7373	0.6907	25	6
VWA (Norway)	19170	0.8013	0.7736	46	11
VWA (Somalia)	1597	0.8167	0.7919	55	11
VWA (Spain)	2500	0.8075	0.7813	33	11

Table S2: Populations showing the highest and lowest standard deviations (σ between parentheses) considering the proportion of hidden mutations across markers, for each familial configuration type. A single-step mutation was simulated in one of the parental meiosis of 1,000,000 configurations of each type, considering the allele frequencies of 10 autosomal STRs in three populations (Norway, Somalia, and Spain) (Kling *et al.*, 2014).

	Duos	**Trios**
Average	σ = 0.124	σ = 0.062
Highest	Norway (σ = 0.142)	Norway (σ = 0.067)
Lowest	Somalia (σ = 0.113)	Somalia (σ = 0.058)

Table S3: Average standard deviation and markers showing the highest and lowest values (in parentheses) of the proportion of hidden mutations across the different population databases: Norway, Somalia, and Spain, for each familial configuration type. A single-step mutation was simulated in one of the parental meiosis of 1,000,000 configurations of each type, considering the allele frequencies of 10 autosomal STRs in three populations (Norway, Somalia, and Spain) (Kling *et al.*, 2014).

	Duos	**Trios**
Average	σ = 0.031	σ = 0.024
Highest	TPOX (σ = 0.061)	TH01 (σ = 0.073)
Lowest	FGA (σ = 0.004)	FGA (σ = 0.003)

2.2.2. X-chromosomal STR mutation rate estimation biases and an informatics correction

Sofia Antão-Sousa [1], Eduardo Conde-Sousa [2], Leonor Gusmão [3], António Amorim [4] and Nádia Pinto [5]

[1] i3S - Health Research and Innovation Institute, Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Faculty of Sciences of University of Porto (FCUP), State University of Rio de Janeiro (UERJ)
[2] i3S - Health Research and Innovation Institute, Instituto de Engenharia Biomédica (INEB)
[3] DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ)
[4] i3S - Health Research and Innovation Institute, Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Biology Department, Faculty of Sciences of University of Porto (FCUP)
[5] i3S - Health Research and Innovation Institute, Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Center of Mathematics of University of Porto (CMUP)

For supplementary material please check Appendix E.

Abstract

Knowledge of the mutational mechanisms of short tandem repeat markers (STRs) is essential for accurately interpreting and quantifying experimental data on a wide range of population genetics problems, specifically in the forensic genetics field. Locus-specific STR mutation rates are usually estimated considering the rates of the observed Mendelian incompatibilities after the genotyping of one-generation relatives, i.e., parent(s)-child duos or trios. This is a faulty approach since, an observed Mendelian incompatibility indicates the occurrence of a mutation, but the latter does not necessarily imply the first. Indeed, underestimations are inevitable when Mendelian incompatibility rates are used as proxies for mutation rates, especially when considering genotyping methodologies of fragment length determination. The likelihood under which one mutation event leads to a Mendelian incompatibility depends on the mode of genetic transmission considered, but also on the parental origin of the mutated allele, the type of 1st degree familial constellation analyzed, and even the genotypes of the analyzed individuals. This implies that the same study design provides distinct levels of confidence for distinct markers and populations; the greater the rate of hidden mutations, the greater the bias in the estimation of mutation rates when Mendelian incompatibility rates are used.

In this work, the biases associated with a typical study design, where Mendelian

incompatibility rates are used as estimates of mutation rates, were weighted, and compared considering diploid and haplodiploid modes of genetic transmission and different 1st degree familial constellations as source data, as well as different markers and populations.

Results showed that biases mainly depend on the mode of genetic transmission and the type of familial data considered. Diploid transmission showed greater biases than haplodiploid, and for both mutation rates estimation biases were generally greater when parent-child duos, instead of trios, were analyzed. The exception was observed for X-chromosomal maternal mutations, where estimates using mother-father-daughter trios were similar to those obtained using mother-son duos.

The wide variation in estimates' accuracy obtained for the different markers, for a specific population and genealogy type, supports a per-marker analysis and prevents gathering data from different genealogical configurations.

Our results support that estimates of per marker haplodiploid mutation rates obtained from the proportion of Mendelian incompatibilities should be estimated by resorting solely to parents-child trios. A framework and corresponding informatics tools to correct Mendelian incompatibilities estimates to accommodate hidden mutations is under construction, extending to the X-linked markers the procedure already developed by us for autosomes.

Keywords: STRs; X chromosome; mutation rates estimation biases; hidden mutations

Introduction

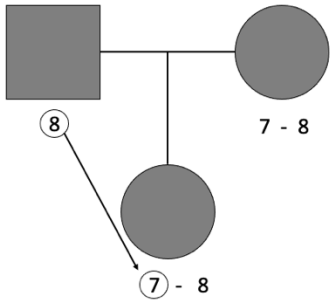
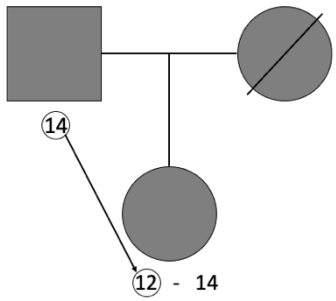
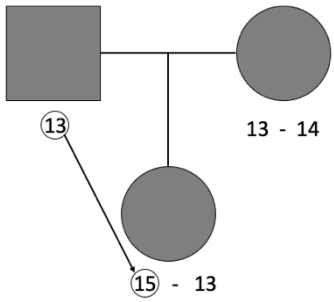
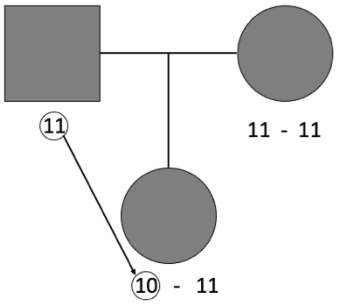
Short tandem repeat markers (STRs) are widely applied in population, evolutionary, and forensic genetics (Cortés-Trujillo *et al.*, 2019; Jobling & Tyler-Smith, 2003; Kayser, 2017; Ren *et al.*, 2018), showing a high level of polymorphism regarding the number of alleles and repeats. This is mainly due to a high rate of length mutations, which results in the number of repeat units differing in the original and the copied allele. Unbiased estimations of STRs mutation rates are, thus, crucial for the proper analysis, interpretation, and quantification of experimental data, as previously claimed by some of us (Amorim & Pinto, 2020; Antão-Sousa *et al.*, 2022b), especially when genotyping methodologies based on fragment length determination are used.

The primary mutational mechanism thought to be leading to changes in STR length is the polymerase template slippage (Schlötterer & Tautz, 1992; Strand *et al.*, 1993). Several factors are known to be correlated with the occurrence of mutations in STRs,

such as the sex and age of the individual (Claerhout *et al.*, 2018), the length of the alleles (Larmuseau & Ottoni, 2018; Schlötterer & Tautz, 1992; Strand *et al.*, 1993; Weber & Wong, 1993; X. Xu *et al.*, 2000), or the sequence of the marker's repetitive motif (Brinkmann *et al.*, 1998; Chakraborty *et al.*, 1997; Dupuy *et al.*, 2004; Edwards *et al.*, 1992; Kruglyak *et al.*, 1998; Valdes *et al.*, 1993). These mutations may occur both somatically and in gametogenesis. In this work, only germinal mutations are considered as these are liable of being transmitted to the offspring. Single-step mutations, i.e., either the gain or loss of a single repeat in the transmitted parental allele, are assumed to be preponderant over those involving multiple repeats, the so-called multistep mutations (Weber & Wong, 1993; X. Xu *et al.*, 2000). This generalized rationale leads to multistep mutations being considered only when a single-step mutation is not able to explain the genotypic configuration observed, which necessarily leads to an overestimation of the single-step mutations compared to multistep ones.

STR mutation rates are usually estimated through the proportioning of Mendelian incompatibilities observed in one-generation pedigrees (duos: either father or mother-child, or trios: both father and mother-child; being father-son links excluded in the case of X-chromosomal markers) - see, for example, (Antão-Sousa *et al.*, 2017; García *et al.*, 2019; Hongdan *et al.*, 2017; Jin *et al.*, 2016; Lan *et al.*, 2018; Liu *et al.*, 2017; Pinto *et al.*, 2020; H. Sun *et al.*, 2014; Zhao *et al.*, 2015). However, length mutations on STR loci do not necessarily lead to Mendelian inconsistencies between the genotypes of parent(s) and child – originating the so-called 'hidden' or 'covert' mutations (Chakraborty *et al.*, 1996; Vicard & Dawid, 2004), as shown in **Figure 1**. The unambiguous identification of which parental allele resulted in which filial one is not attainable for most of the parent(s)-child genotypic configurations, for diploid and haplodiploid systems. Haploid systems, as the non-recombining region of the Y chromosome, have been presented as the only case where length mutations are unambiguously detected (Pinto *et al.*, 2014a). Notwithstanding, it should be noted that even in that case hidden mutations may occur when considering markers with complex structures, as previously shown (Dupuy *et al.*, 2004).

Figure 1: Examples of allele transmission interpretation when the most parsimonious reasoning (mutation involving the least number of steps possible) is considered, for X-STRs, familial duos and trios. The arrow and circles indicate the parent-child allele transmission.

Representation of the familial constellation and mutation occurred	**Familial constellation considered**	**Interpretation under the most parsimonious reasoning**
a. 	Father-mother-daughter trio	Transmission of paternal allele 8 and maternal allele 7.
b. 	Father-daughter duo	Transmission of paternal allele 14.
c. 	Father-mother-daughter trio	Transmission of paternal allele 13 and maternal transmission of mutated allele 14.
d. 	Father-mother-daughter trio	Ambiguous parental origin of mutation.

The naïf approach of counting observed Mendelian incompatibilities leads, necessarily, to underestimated mutation rates. Moreover, for diploid and haplodiploid loci, the presence of silent alleles may also distort these estimates. However, with the genotyping methodologies currently employed, their presence can be ruled out if precaution is taken when analyzing experimental results, hence they will not be considered in this work.

The generalized approach of estimating mutation rates through incompatibility observation at diploid and haplodiploid loci is thus biased. These biases are already known to be greater when duos, instead of trios, are analyzed, for both autosomal and X-chromosomal markers (Antão-Sousa, Conde-Sousa, *et al.*, 2019; Antão-Sousa *et al.*, 2022b; Chakraborty *et al.*, 1996; Slooten & Ricciardi, 2013; Vicard *et al.*, 2008; Vicard & Dawid, 2004). The impact that these biases may have on distinct modes of genetic transmission, markers, and populations was however never compared.

Mathematical formulae to correct locus-specific incompatibility rates were already presented for autosomal STRs (Antão-Sousa *et al.*, 2022b; Chakraborty *et al.*, 1996; Slooten & Ricciardi, 2013; Vicard *et al.*, 2008; Vicard & Dawid, 2004), most of them assuming a specific mutation model, perhaps incurring in a circular reasoning (Chakraborty *et al.*, 1996; Vicard *et al.*, 2008; Vicard & Dawid, 2004). The simple, uncorrected, Mendelian incompatibilities counting method is still the generally accepted approach to estimating mutation rates - see (García *et al.*, 2019; Hongdan *et al.*, 2017; Jin *et al.*, 2016; Lan *et al.*, 2018; Liu *et al.*, 2017; H. Sun *et al.*, 2014; Zhao *et al.*, 2015), for example. Indeed, in the literature review we performed for this work, no studies were found where any kind of correction was performed and in all the cases the mutation rates were estimated directly from Mendelian incompatibility rates. This can be due to only recently an informatics tool was provided to compute such correction (Antão-Sousa *et al.*, 2022b).

In this work, we quantified the biases inherent to the estimation of mutation rates considering the simple counting of Mendelian incompatibilities for haplodiploid modes of transmission, for distinct familial constellations, markers, and populations. We aimed to measure and compare the levels of confidence reached in each case, and also to compare with the one reached for diploid transmission. All the analyses were performed resorting to simulated familial genotypic constellations, using allele frequencies from real population databases, and considering single-, two-, and three-step mutations. X-linked transmission modes were considered, and the biases were quantified and compared taking into account (a) the population allele frequencies distribution, (b) the type of experimental data (family duos or trios, varying the sex of the individuals), but also (c) the mode of genetic transmission, and (d) the type of mutation that occurred.

Material and methods

Familial genotypic configurations were generated by resorting to Python™ programming language, considering parent(s)-child duos and trios for autosomal and X-chromosomal mode of genetic transmission. For X-chromosomal markers, father-mother-daughter trios, and mother-daughter, father-daughter, and mother-son duos were simulated and for autosomal markers, father-mother-child trios, and parent-child duos. Parental genotypes were randomly attributed, considering the population allele frequency distribution.

Duos and trios' genotypic configurations were generated considering allele frequencies for the 12 X-STRs comprised in the Investigator Argus X-12 QS kit for Argentinean, Eastern Asian, and Northern European populations (Kling *et al.*, 2015). For each marker, type of pedigree, and population database, 1,000,000 familial genotypic constellations were simulated. In each simulated familial genotypic constellation one single, one two- or one three-step mutation was simulated in one of the transmitted parental alleles. When trios were considered, paternal and maternal mutations were both simulated, one at a time, and were independently analyzed. Mutated alleles were assumed to be transmitted to the offspring, while the other filial allele (when present) was randomly selected from either the population (in the case of duos) or the other parent (in the case of trios). For each type of pedigree considered, 10^6 mutations were simulated, and the outcome in terms of compatibility/incompatibility, most parsimonious parental origin (paternal/maternal), and type of mutation (single/multistep) was assessed and compared.

To summarize, for the haplodiploid mode of genetic transmission, one single-, one two-, and one three-step mutation was simulated for each familial genotypic configuration, assuming each one of the 1st degree familial constellations, markers, and population databases previously mentioned. This resulted in the analysis of 5.4×10^8 X-linked genotypic constellations = 5 (types of familial constellations: father-mother-daughter trios, paternal and maternal mutation, and mother-daughter, father-daughter, and mother-son duos) × 3 (single-, two-, and three-step mutations) × 12 (X-STRs) × 3 (population databases) × 10^6 (familial genotype simulations).

Each simulation resulted in one of two possible outcomes: the family genotypes were incompatible with codominant Mendelian inheritance, or otherwise. Under the standard, general practice, the rate of the cases resulting in familial genotypic incompatibility for a specific marker would be presented as the marker-specific average mutation rate; while the rate of cases that resulted in familial genotypic compatibility equates to the rate of

hidden mutations. The greater the rate of hidden mutations, the greater the bias of the mutation rates estimates obtained through the analysis of Mendelian incompatibilities.
This analysis allowed the comparison of the biases obtained when different study designs, markers, or populations are considered. The autosomal mode of genetic transmission was also considered to compare biases associated with the mode of genetic transmission and not allele frequencies.
Fisher Exact tests to ascertain p values were computed considering a level of significance equal to 0.05. Python algorithms used for the presented simulations will be publicly and freely available online. An informatics tool that allows not only the replication of our results but also the correction of Mendelian incompatibility rates into mutation ones, for any marker, population, and familial configuration, assuming single-step mutations is under construction. This tool and corresponding algorithms will also be publicly and freely available online.

Biases in X-chromosomal mutation rates estimation

Hidden mutations

In this section, the results comparing X-STR mutation and incompatibility rates are presented considering different markers and populations and all the possible 1st degree familial configurations.

Considering the familial pedigree

The average rate of hidden mutations obtained for the different familial data was computed - see **Table 1**.
The accuracy of the estimates showed to be heavily dependent on parental and offspring sexes within the familial data analyzed. For example, mother-daughter duos hid, on average, 2.8 and 2.1 times more single-step mutations than mother-son and father-daughter duos, respectively.
Moreover, the accuracy of the estimates depends not only on the type of duos analyzed but also, in the case of trios, on which parental meiosis (either maternal or paternal) the mutation occurred. Indeed, when trios were analyzed, maternal single-step mutations were detected 4.2 times more often than the paternal ones.
Considering paternal single-step mutations specifically, these were hidden 5.6 times more often in father-daughter duos than in trios. In the case of maternal mutations, similar estimates were obtained in mother-son duos and mother-father-daughter trios

(14.4%), contrarily to mother-daughter duos where 2.8 more hidden mutations were found (40.5%).

Table 1: Average rates of hidden mutations when one single-, one two-, or one three-steps mutation was simulated in mother-daughter, mother-son, father-daughter duos, and in trios with daughters (mutation simulated in paternal and maternal meiosis).

	Average rate of hidden mutations				
	Duos			Trios (with daughters)	
Type of mutation	Mother-Daughter	Mother-Son	Father-Daughter	Meiosis in which mutation was simulated	
				Paternal	Maternal
1 step	0.405	0.144	0.191	0.034	0.144
2 steps	0.385	0.089	0.191	0.016	0.089
3 steps	0.363	0.049	0.191	0.007	0.049

One single-, one two-, and one three-steps mutation simulated in a randomly selected parental meiosis of 1,000,000 parent-child duos and parents-daughter trios, considering the allele frequencies of 12 X-STRs in the populations of Argentina, Eastern Asia, and Northern Europe (Kling *et al.*, 2015). N = 36,000,000, for each family and mutation type.

Considering the mode of transmission

The accuracy of the mutation estimates obtained through the rate of Mendelian incompatibilities was compared considering the autosomal and the X-chromosomal modes of genetic transmission. To circumvent the influence of the allele distribution of the different markers, the same allele frequency database was used, considering either autosomal or X-chromosomal modes of transmission – see **Table 2**.

Table 2: Average proportion of hidden mutations for autosomal and X-chromosomal modes of genetic transmission, when the same allele frequency database is used for the simulations and corresponding analyses. Different familial configurations were also considered.

Single-step mutation simulated	Autosomal mode of transmission	Parent-child duos			Parents-child trios	
		0.409			0.165	
	X-chromosomal mode of transmission	Mother-daughter duos	Mother-son duos	Father-daughter duos	Trios with mutation in paternal meiosis	Trios with mutation in maternal meiosis
		0.409	0.131	0.192	0.028	0.131
Two- step mutation simulated	Autosomal mode of transmission	Parent-child duos			Parents-child trios	
		0.388			0.111	
	X-chromosomal mode of transmission	Mother-daughter duos	Mother-son duos	Father-daughter duos	Trios with mutation in paternal meiosis	Trios with mutation in maternal meiosis
		0.388	0.087	0.192	0.017	0.088
Three-step mutation simulated	Autosomal mode of transmission	Parent-child duos			Parents-child trios	
		0.369			0.074	
	X-chromosomal mode of transmission	Mother-daughter duos	Mother-son duos	Father-daughter duos	Trios with mutation in paternal meiosis	Trios with mutation in maternal meiosis
		0.369	0.059	0.192	0.012	0.059

One single-, one two-, and one three-steps mutations simulated in a parental meiosis of 1,000,000 family configurations of each type, considering a single database of 10 STRs in three populations (Kling *et al.*, 2014).

N = 30,000,000, for each family and mutation type.

As expected, the rate of hidden mutations was the highest when parent-child duos were used to estimate autosomal mutation rates, and (equivalently) when mother-daughter duos were considered for X-chromosomal transmission. The lowest proportion of hidden mutations was found when trios were analyzed, specifically when a paternal mutation at an X-linked marker was simulated. Generally, a smaller proportion of hidden mutations is expected when X chromosome transmission is analyzed. For example, a paternal autosomal single-step mutation was not detected in 40.9% of the cases, the same

probability was obtained for maternal mutations when mother-daughter duos were analyzed in X chromosome markers. This probability declined more than a half when mother-son (13.1%) or father-daughter (19.2%) duos were analyzed. In the case of trios, paternal and maternal X-STR single-step mutations were detected 5.9 and 1.3 times more often than the autosomal ones.

Considering the marker analyzed

The markers for which the best and worst mutation rate estimates were obtained through the analysis of the resulting mendelian incompatibilities were DXS10135, and DXS8378 and DXS7423, respectively – see **Table S1**. Indeed, DXS10135 showed consistently the lowest rates of hidden single-step mutations for all the familial pedigrees and populations (ranging from 0.004 – paternal mutation analyzed in trios for Northern Europe population, to 0.197 – maternal mutation analyzed in mother-daughter duos for Eastern Asian population). The highest rates of hidden mutations were found for either DXS8378 or DXS7423 (ranging from 0.692 - DXS7423, maternal mutation analyzed in mother-daughter duos for Eastern Asian population, to 0.068 – DXS7423, paternal mutation analyzed in trios for Northern European population).

The X-chromosomal markers revealed diverse behaviors regarding the detection of mutations depending on the population and type of genealogy analyzed, most showing statistically significant differences – see **supplementary material Appendix 1**.

The standard deviation concerning the proportion of hidden single-step mutations across the different markers varied from 0.0219 (for the Argentinean population, in trios with paternal mutation) to 0.1643 (for Eastern Asian population, in mother-daughter duos), showing that within a population the biases associated with the different markers' mutation rate estimates vary substantially. Globally, the greatest values of standard deviation were found when considering mother-daughter duos – see **Table S2**.

Considering the populations analyzed

Some populations showed, broadly speaking, better estimates (i.e., lower rates of hidden mutations) than others – see **Table S3**. Globally, when single-step mutations were considered, Argentina showed the best estimates, as lower rates of hidden mutations were found. Conversely, Eastern Asia showed the worst estimates, with the highest proportions of hidden mutations.

Also, statistically significant differences were found among the populations under study in what concerns the rate of hidden mutations in the different markers – see **supplementary material Appendix 1**. For example, when the differences between the incompatibility and hidden mutation rates were compared using mother-son duos for the same marker and considering different populations, for single-step mutations, only in two cases the p-values showed to be greater than 0.00001: Asia and Northern Europe for DXS10079 ($p = 0.326$) and DXS10101 ($p = 0.0193$). The greatest average standard deviation between populations was found for mother-daughter duos (0.039); trios and mother-son duos showed the lowest values – see **Table S4**.

Discussion and conclusion

Mutation rates are generally estimated through the proportion of Mendelian incompatibilities observed in parent(s)-child duos or trios genotyped through methodologies of fragment length determination. The accuracy of these estimates varies between markers and populations (depending on the allelic frequency distributions), but - more importantly, with the type of family data (parent(s)-child duos or trios) and with the mode of transmission. In autosomal transmission, parental and filial alleles cannot be undoubtedly identified, and four allele transitions possibilities can be established when duos are analyzed (eight in case of trios), even assuming no double mutation (both paternal and maternal) and the absence of silent alleles. The identification of which allelic transfer broke the identity by descent cannot be guaranteed, but only assumed. In X-chromosomal transmission, a less ambiguous situation is found, as male hemizygosity allows the unambiguous identification of either the original (parental) allele in father-daughter links or the transmitted (filial) allele in mother-son ones.

In both diploid and haplodiploid systems, mutations do not necessarily lead to Mendelian incompatibilities, and therefore hidden mutations may occur. Mutation rates are necessarily underestimated if the method of the simple counting of Mendelian inconsistencies observed is used. Hence, in this work and a previous one (Antão-Sousa *et al.*, 2022b), we focused on these systems as they harbor the most biases.

In most cases, biases are more important when duos, instead of trios, are used. Previous work clearly shows that autosomal Mendelian incompatibility rates in duos provide less accurate estimates for mutation rates than trios (Antão-Sousa *et al.*, 2022b; Slooten & Ricciardi, 2013), which prevents the simple pooling of the two types of data. Nonetheless, 32 from 60 papers found in the PubMed® database (March 1st, 2022) reporting autosomal

or X-chromosomal STR mutation rates joined data from duos and trios for these estimates.

In the case of X-chromosomal transmission, the information retrieved from trios is also more robust than that obtained from duos, despite similar estimates obtained when mother-son duos and father-mother-daughter trios were analyzed for maternal mutations.

Estimates obtained through the analysis of Mendelian incompatibilities are generally more accurate for X-chromosome markers than for autosomes, which means that, under similar conditions, the reported overall mutation rate estimates for X-STRs are more precise than those for autosomes. Indeed, the accuracy of the autosomal average mutation estimates when trios are analyzed is similar to the one obtained for X chromosome markers when mother-son or father-daughter duos are considered. Inversely, autosomes provide a more accurate estimation of the ratio between maternal and paternal mutations than the X-chromosome; the latter showing a bias favoring the identification of paternal mutations compared with maternal ones. All in all, as maternal and paternal mutations are usually both of interest, the use of parent-child duos to estimate average mutation rates for diploid and haplodiploid modes of genetic transmission is strongly discouraged due to the high rate of expected hidden mutations. The probability of occurrence of hidden mutations depends on the allele frequency distribution and, therefore, under the classic Mendelian incompatibilities approach, some markers or populations are expected to have more accurate estimates than others. For example, for X-chromosomal transmission in father-daughter duos, any occurring mutation does not result in Mendelian incompatibility whenever the maternal allele received by the daughter is the same as the father's allele. This occurs with a probability equal to the expected homozygosity of the marker: $\sum_a f_a^2$, where f_a is the frequency of the allele *a*. Remarkably, in father-daughter duos, the probability of a paternal mutation being compatible with Mendelian X-chromosomal transmission is independent of the mutation type and linearly correlated with the expected homozygosity of the marker. This singular situation is not shared by any other duos or trios, whatever the transmission mode (X-linked or autosomal) – see **Appendix 2** for algebraic formulae.

Since STRs included in forensic kits are usually selected among those with a higher diversity within rather than between populations, a slight variation of biases can be expected for other populations not included in this study. On the other hand, a slight variation in the proportion of marker-specific mutation biases in different populations was observed, mainly when trios were analyzed. Since rare events are at stake, it is important to use the largest data collection possible to obtain accurate mutation estimates. Hence,

despite the significant differences found in most pairwise analyses computed between populations, the small amount of variation observed supports the strategy of pooling data from different populations, geographically close and after computing differentiation analyses, as long as the same genealogies are considered. On the other hand, the different rates of hidden mutations associated with the different markers support a mutation estimation and correction specifically devoted to each marker. Per locus analysis demand, however, that mutation data must be made available in a format allowing the pooling from different laboratories, as well as the development and enhancement of mathematical models considering other factors known to influence mutation, as is the case of the parental allele length in terms of repeat number (Amorim & Pinto, 2020).

The framework we present in this work can be used to correct the mutation rates estimated through the Mendelian incompatibilities approach. To obtain the proposed corrective factor, it is just necessary to know the distribution of the allele frequencies at the loci of interest and to perform the same simulations here described. The informatics tool which will allow these computations is under construction and will be made publicly available as soon as possible. As an example, if parent-child duos are used and a rate R of Mendelian incompatibilities is found at the DXS8378 locus in the Argentinian population for mother-daughter duos, the corrected value of R/(1-0.613) should be used as the estimated one-step mutation rate for this locus and population.

Appendix 2

The frequency of hidden mutations depends on the type of family configuration and allele frequencies distribution of the population analyzed. Algebraic formulae for the genotypic configurations able to hide mutations are presented below for both autosomal and X-chromosomal transmission, considering the analysis of duos for simplicity's sake. Also, the population in Hardy-Weinberg equilibrium and the absence of silent alleles were assumed.

The simplest case corresponds to father-daughter duos and X-chromosome transmission, as all mutations are covert whenever the maternally transmitted allele is the same as the father's allele. This occurs with probability $\sum_a f_a^2$, where f_a is the frequency of the allele *a*. Indeed, the probability of hidden mutations occurring when an X-specific marker is analyzed in father-daughter duos is independent of the type of mutation and linearly correlated with the expected homozygosity of the marker in the population.

When mother-son duos are analyzed for X-chromosome markers, mutations are covert whenever the mother is heterozygous exhibiting the mutated allele. Thus, a k-step mutation, $k \in \mathbb{N}$, may be covert with a probability equal to $\sum_a f_a(f_{a-k} + f_{a+k})$, where f_i is the frequency of the allele *i*. Contrarily to the previous duo type, when mother-son duos are analyzed, the probability of mutations resulting in hidden ones, not only depends on the type of the mutation, as it is inversely correlated with the expected homozygosity of the marker.

Finally, when mother-daughter or parent-child (sex not discriminated) duos are analyzed for X-chromosomal or autosomal markers, respectively, mutations may be hidden in both cases: when the parent is heterozygous showing the mutated allele, or when the allele received from the non-typed parent is identical to the mutated one. Thus, a k-step mutation, $k \in \mathbb{N}$, may be hidden with a probability equal to $\sum_a f_a\left(2f_{a-k} + 2f_{a+k} - f_{a-k}^2 - f_{a+k}^2\right)$, where f_i is the frequency of the allele *i*.

Acknowledgments

This work was partially financed by FEDER—Fundo Europeu de Desenvolvimento Regional - funds through the COMPETE 2020—Operacional Programme for Competitiveness and Internationalisation (POCI), Portugal 2020; Portuguese funds through FCT—Fundação para a Ciência e a Tecnologia/Ministério da Ciência, Tecnologia e Inovação in the framework of the Decreto-Lei nº 57/2016 de 29 de Agosto (NP work contract) and through the doctoral grant SFRH/BD/136284/2018; and the projects "Institute for Research and Innovation in Health Sciences" (POCI-01-0145-FEDER-007274) and "Center of Mathematics of the University of Porto" (UID/MAT/00144/2019); FCT National Roadmap of Research Infrastructures project "PPBI: PPBI - Portuguese Platform of BioImaging" PPBI-POCI-01-0145-FEDER-022122. LG (ref. 305330/2016-0) was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq and Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro – FAPERJ (CNE-2022).

References

[1] D. Kling, A. O. Tillmar, and T. Egeland, "Familias 3 – Extensions and new functionality," *Forensic Sci Int Genet*, vol. 13, pp. 121–127, Nov. 2014, doi: 10.1016/j.fsigen.2014.07.004.

[2] M. Kayser, "Forensic use of Y-chromosome DNA: a general overview," *Hum Genet*, vol. 136, no. 5, p. 621, May 2017, doi: 10.1007/S00439-017-1776-9.

[3] M. A. Jobling and C. Tyler-Smith, "The human Y chromosome: an evolutionary marker comes of age," *Nature Reviews Genetics 2003 4:8*, vol. 4, no. 8, pp. 598–612, Aug. 2003, doi: 10.1038/nrg1124.

[4] I. Cortés-Trujillo *et al.*, "Allele and haplotype frequencies of 12 X-STRs in Mexican population," *Forensic Sci Int Genet*, vol. 38, pp. e11–e13, Jan. 2019, doi: 10.1016/J.FSIGEN.2018.10.012.

[5] Z. Ren *et al.*, "Population genetic data of 22 autosomal STRs in the Guizhou Miao population, southwestern China," *Forensic Sci Int Genet*, vol. 32, pp. e7–e8, Jan. 2018, doi: 10.1016/J.FSIGEN.2017.10.007.

[6] A. Amorim and N. Pinto, "Estimates of mutation rates from incompatibilities are misleading - guidelines for publication and retrieval of mutation data urgently needed," 2020, doi: 10.1016/j.fsigss.2019.10.110.

[7] S. Antão-Sousa, E. Conde-Sousa, L. Gusmão, A. Amorim, and N. Pinto, "Estimations of Mutation Rates Depend on Population Allele Frequency Distribution: The Case of Autosomal Microsatellites," *Genes (Basel)*, vol. 13, no. 7, Jul. 2022, doi: 10.3390/genes13071248.

[8] M. Strand, T. A. Prolla, R. M. Liskay, and T. D. Petes, "Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair," *Nature 1993 365:6443*, vol. 365, no. 6443, pp. 274–276, 1993, doi: 10.1038/365274a0.

[9] C. Schlötterer and D. Tautz, "Slippage synthesis of simple sequence DNA.," *Nucleic Acids Res*, vol. 20, no. 2, p. 211, Jan. 1992, doi: 10.1093/NAR/20.2.211.

[10] S. Claerhout *et al.*, "Determining Y-STR mutation rates in deep-routing genealogies: Identification of haplogroup differences," *Forensic Sci Int Genet*, vol. 34, pp. 1–10, May 2018, doi: 10.1016/J.FSIGEN.2018.01.005.

[11] X. Xu, M. Peng, Z. Fang, and X. Xu, "The direction of microsatellite mutations is dependent upon allele length," *Nat Genet*, vol. 24, no. 4, pp. 396–399, Apr. 2000, doi: 10.1038/74238.

[12] J. L. Weber and C. Wong, "Mutation of human short tandem repeats," *Hum Mol Genet*, vol. 2, no. 8, pp. 1123–1128, Aug. 1993, doi: 10.1093/HMG/2.8.1123.

[13] M. H. D. Larmuseau and C. Ottoni, "Mediterranean Y-chromosome 2.0-why the Y in the Mediterranean is still relevant in the postgenomic era," *Ann Hum Biol*, vol. 45, no. 1, pp. 20–33, Jan. 2018, doi: 10.1080/03014460.2017.1402956.

[14] B. Brinkmann, M. Klintschar, F. Neuhuber, J. Hühne, and B. Rolf, "Mutation Rate in Human Microsatellites: Influence of the Structure and Length of the Tandem Repeat," *The American Journal of Human Genetics*, vol. 62, no. 6, pp. 1408–1415, Jun. 1998, doi: 10.1086/301869.

[15] A. M. Valdes, M. Slatkin, and N. B. Freimer, "Allele frequencies at microsatellite loci: the stepwise mutation model revisited," *Genetics*, vol. 133, no. 3, pp. 737–749, 1993, doi: 10.1093/GENETICS/133.3.737.

[16] R. Chakraborty, M. Kimmel, D. N. Stivers, L. J. Davison, and R. Deka, "Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci," *Proc Natl Acad Sci U S A*, vol. 94, no. 3, pp. 1041–1046, Feb. 1997, doi: 10.1073/PNAS.94.3.1041.

[17] A. Edwards, H. A. Hammond, L. Jin, C. T. Caskey, and R. Chakraborty, "Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups," *Genomics*, vol. 12, no. 2, pp. 241–253, 1992, doi: 10.1016/0888-7543(92)90371-X.

[18] S. Kruglyak, R. T. Durrett, M. D. Schug, and C. F. Aquadro, "Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations," *Proc Natl Acad Sci U S A*, vol. 95, no. 18, pp. 10774–10778, Sep. 1998, doi: 10.1073/PNAS.95.18.10774.

[19] B. M. Dupuy, M. Stenersen, T. Egeland, and B. Olaisen, "Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci," *Hum Mutat*, vol. 23, no. 2, pp. 117–124, 2004, doi: 10.1002/HUMU.10294.

[20] B. Jin *et al.*, "Mutational analysis of 33 autosomal short tandem repeat (STR) loci in southwest Chinese Han population based on trio parentage testing," *Forensic Sci Int Genet*, vol. 23, pp. 86–90, Jul. 2016, doi: 10.1016/J.FSIGEN.2016.03.009.

[21] N. Pinto *et al.*, "Paternal and maternal mutations in X-STRs: A GHEP-ISFG collaborative study," *Forensic Sci Int Genet*, vol. 46, May 2020, doi: 10.1016/J.FSIGEN.2020.102258.

[22] M. G. García, C. I. Catanesi, G. A. Penacino, L. Gusmão, and N. Pinto, "X-chromosome data for 12 STRs: Towards an Argentinian database of forensic haplotype frequencies," *Forensic Sci Int Genet*, vol. 41, pp. e8–e13, Jul. 2019, doi: 10.1016/J.FSIGEN.2019.04.005.

[23] S. Antão-Sousa *et al.*, "Mutation rates and segregation data on 16 Y-STRs: An update to previous GHEP-ISFG studies," *Forensic Sci Int Genet Suppl Ser*, vol. 6, pp. e601–e602, Dec. 2017, doi: 10.1016/J.FSIGSS.2017.10.008.

[24] H. Sun, S. Liu, Y. Zhang, and M. R. Whittle, "Comparison of southern Chinese Han and Brazilian Caucasian mutation rates at autosomal short tandem repeat loci used in human forensic genetics," *Int J Legal Med*, vol. 128, no. 1, pp. 1–9, Jan. 2014, doi: 10.1007/S00414-013-0847-2.

[25] Z. Zhao *et al.*, "Mutation rate estimation for 15 autosomal STR loci in a large population from Mainland China," *Meta Gene*, vol. 5, pp. 150–156, Sep. 2015, doi: 10.1016/J.MGENE.2015.07.006.

[26] W. Hongdan *et al.*, "Evaluation of the genetic parameters and mutation analysis of 22 STR loci in the central Chinese Han population," *Int J Legal Med*, vol. 131, no. 1, pp. 103–105, Jan. 2017, doi: 10.1007/S00414-016-1389-1.

[27] Q. L. Liu, Y. F. Chen, X. L. Huang, K. Y. Liu, H. Zhao, and D. J. Lu, "Population data and mutation rates of 19 STR loci in seven provinces from China based on Goldeneye™ DNA ID System 20A," *Int J Legal Med*, vol. 131, no. 3, pp. 653–656, May 2017, doi: 10.1007/S00414-016-1441-1.

[28] Q. Lan *et al.*, "Mutability analysis towards 21 STR loci included in the AGCU 21 + 1 kit in Chinese Han population," *Int J Legal Med*, vol. 132, no. 5, pp. 1287–1291, Sep. 2018, doi: 10.1007/S00414-018-1873-X.

[29] P. Vicard and A. P. Dawid, "A statistical treatment of biases affecting the estimation of mutation rates," *Mutat Res*, vol. 547, no. 1–2, pp. 19–33, Mar. 2004, doi: 10.1016/J.MRFMMM.2003.11.005.

[30] R. Chakraborty, D. N. Stivers, and Y. Zhong, "Estimation of mutation rates from parentage exclusion data: applications to STR and VNTR loci," *Mutat Res*, vol. 354, no. 1, pp. 41–48, Jul. 1996, doi: 10.1016/0027-5107(96)00014-0.

[31] N. Pinto, L. Gusmão, and A. Amorim, "Mutation and mutation rates at Y chromosome specific Short Tandem Repeat Polymorphisms (STRs): a reappraisal," *Forensic Sci Int Genet*, vol. 9, no. 1, pp. 20–24, Mar. 2014, doi: 10.1016/J.FSIGEN.2013.10.008.

[32] P. Vicard, A. P. Dawid, J. Mortera, and S. L. Lauritzen, "Estimating mutation rates from paternity casework," *Forensic Sci Int Genet*, vol. 2, no. 1, pp. 9–18, Jan. 2008, doi: 10.1016/J.FSIGEN.2007.07.002.

[33] K. Slooten and F. Ricciardi, "Estimation of mutation probabilities for autosomal STR markers," *Forensic Sci Int Genet*, vol. 7, no. 3, pp. 337–344, May 2013, doi: 10.1016/J.FSIGEN.2013.01.006.

[34] S. Antão-Sousa, E. Conde-Sousa, L. Gusmão, A. Amorim, and N. Pinto, "Underestimation and misclassification of mutations at X chromosome STRs depend on population's allelic profile," *Forensic Sci Int Genet Suppl Ser*, vol. 7, no. 1, pp. 718–720, Dec. 2019, doi: 10.1016/J.FSIGSS.2019.10.150.

[35] D. Kling, A. Tillmar, T. Egeland, and P. Mostad, "A general model for likelihood computations of genetic marker data accounting for linkage, linkage disequilibrium, and mutations," *Int J Legal Med*, vol. 129, no. 5, pp. 943–954, Sep. 2015, doi: 10.1007/S00414-014-1117-7.

[36] H. Ellegren, "Microsatellites: simple sequences with complex evolution," *Nature Reviews Genetics 2004 5:6*, vol. 5, no. 6, pp. 435–445, Jun. 2004, doi: 10.1038/nrg1348.

[37] S. Panzer, D. P. A. Kuhl, and C. T. Caskey, "Unstable triplet repeat sequences: A source of cancer mutations?," *Stem Cells*, vol. 13, no. 2, pp. 146–157, Jan. 1995, doi: 10.1002/STEM.5530130206.

[38] C. E. Pearson, K. N. Edamura, and J. D. Cleary, "Repeat instability: mechanisms of dynamic mutations," *Nature Reviews Genetics 2005 6:10*, vol. 6, no. 10, pp. 729–742, Oct. 2005, doi: 10.1038/nrg1689.

[39] K. A. Eckert and S. E. Hile, "Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome," *Mol Carcinog*, vol. 48, no. 4, pp. 379–388, Apr. 2009, doi: 10.1002/MC.20499.

[40] T. Ohta and M. Kimura, "A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population," *Genet Res*, vol. 22, no. 2, pp. 201–204, 1973, doi: 10.1017/S0016672300012994.

[41] C. F. Wehrhahn, "The evolution of selectively similar electrophoretically detectable alleles in finite natural populations," *Genetics*, vol. 80, no. 2, pp. 375–394, 1975, doi: 10.1093/GENETICS/80.2.375.

[42] M. Gymrek, T. Willems, D. Reich, and Y. Erlich, "Interpreting short tandem repeat variations in humans using mutational constraint," *Nature Genetics 2017 49:10*, vol. 49, no. 10, pp. 1495–1501, Sep. 2017, doi: 10.1038/ng.3952.

[43] J. X. Sun *et al.*, "A direct characterization of human mutation based on microsatellites," *Nat Genet*, vol. 44, no. 10, pp. 1161–1165, Oct. 2012, doi: 10.1038/NG.2398.

[44] H. Xu and Y. X. Fu, "Estimating Effective Population Size or Mutation Rate With Microsatellites," *Genetics*, vol. 166, no. 1, pp. 555–563, Jan. 2004, doi: 10.1534/GENETICS.166.1.555.

[45] R. Z. Aandahl, J. F. Reyes, S. A. Sisson, and M. M. Tanaka, "A Model-Based Bayesian Estimation of the Rate of Evolution of VNTR Loci in Mycobacterium tuberculosis," *PLoS Comput Biol*, vol. 8, no. 6, p. e1002573, Jun. 2012, doi: 10.1371/JOURNAL.PCBI.1002573.

[46] S. Barthe, F. Gugerli, N. A. Barkley, L. Maggia, C. Cardi, and I. Scotti, "Always Look on Both Sides: Phylogenetic Information Conveyed by Simple Sequence Repeat Allele Sequences," *PLoS One*, vol. 7, no. 7, p. e40699, Jul. 2012, doi: 10.1371/JOURNAL.PONE.0040699.

[47] "Forensic Mathematics." https://dna-view.com/#Software (accessed Apr. 28, 2022).

[48] N. Pinto, L. Gusmão, and A. Amorim, "Mutation and mutation rates at y chromosome specific Short Tandem Repeat Polymorphisms (STRs): A reappraisal," *Forensic Sci Int Genet*, vol. 9, no. 1, pp. 20–24, Mar. 2014, doi: 10.1016/j.fsigen.2013.10.008.

[49] C. H. Brenner, "Multiple mutations, covert mutations and false exclusions in paternity casework," *Int Congr Ser*, vol. 1261, no. C, pp. 112–114, Apr. 2004, doi: 10.1016/S0531-5131(03)01843-0.

[50] R. Lessig and J. Edelmann, "Y chromosome polymorphisms and haplotypes in west Saxony (Germany)," *Int J Legal Med*, vol. 111, no. 4, pp. 215–218, Jun. 1998, doi: 10.1007/S004140050155.

[51] N. O. Blanchi *et al.*, "Characterization of Ancestral and Derived Y-Chromosome Haplotypes of New World Native Populations," *The American Journal of Human Genetics*, vol. 63, no. 6, pp. 1862–1871, Dec. 1998, doi: 10.1086/302141.

[52] M. Kayser *et al.*, "Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs.," *Am J Hum Genet*, vol. 66, no. 5, p. 1580, 2000, doi: 10.1086/302905.

[53] B. M. Dupuy, M. Stenersen, T. Egeland, and B. Olaisen, "Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci," *Hum Mutat*, vol. 23, no. 2, pp. 117–124, 2004, doi: 10.1002/HUMU.10294.

[54] L. C. Tsai *et al.*, "Haplotype frequencies of nine Y-chromosome STR loci in the Taiwanese Han population," *Int J Legal Med*, vol. 116, no. 3, pp. 179–183, 2002, doi: 10.1007/S004140100236.

[55] R. Kurihara *et al.*, "Mutations in 14 Y-STR loci among Japanese father-son haplotypes," *Int J Legal Med*, vol. 118, no. 3, pp. 125–131, Jun. 2004, doi: 10.1007/S00414-003-0422-3.

[56] B. Budowle *et al.*, "Twelve short tandem repeat loci Y chromosome haplotypes: genetic analysis on populations residing in North America," *Forensic Sci Int*, vol. 150, no. 1, pp. 1–15, May 2005, doi: 10.1016/J.FORSCIINT.2005.01.010.

[57] B. Berger, A. Lindinger, H. Niederstätter, P. Grubwieser, and W. Parson, "Y-STR typing of an Austrian population sample using a 17-loci multiplex PCR assay," *Int J Legal Med*, vol. 119, no. 4, pp. 241–246, 2005, doi: 10.1007/S00414-005-0546-8.

[58] L. Gusmão *et al.*, "Mutation rates at Y chromosome specific microsatellites," *Hum Mutat*, vol. 26, no. 6, pp. 520–528, Dec. 2005, doi: 10.1002/HUMU.20254.

[59] D. J. Ballard *et al.*, "A study of mutation rates and the characterisation of intermediate, null and duplicated alleles for 13 Y chromosome STRs," *Forensic Sci Int*, vol. 155, no. 1, pp. 65–70, Dec. 2005, doi: 10.1016/J.FORSCIINT.2004.12.012.

[60] S. Turrina, R. Atzei, and D. de Leo, "Y-chromosomal STR haplotypes in a Northeast Italian population sample using 17plex loci PCR assay," *Int J Legal Med*, vol. 120, no. 1, pp. 56–59, Jan. 2006, doi: 10.1007/S00414-005-0054-X.

[61] H. Y. Lee *et al.*, "Haplotypes and mutation analysis of 22 Y-chromosomal STRs in Korean father-son pairs," *Int J Legal Med*, vol. 121, no. 2, pp. 128–135, Mar. 2007, doi: 10.1007/S00414-006-0130-X.

[62] C. Hohoff, K. Dewa, U. Sibbing, K. Hoppe, P. Forster, and B. Brinkmann, "Y-chromosomal microsatellite mutation rates in a population sample from northwestern

Germany," *Int J Legal Med*, vol. 121, no. 5, pp. 359–363, Sep. 2007, doi: 10.1007/S00414-006-0123-9.

[63] R. Kumagai, A. Kumagai, K. Saigusa, and Y. Aoki, "Haplotype analysis of 17 Y-STR loci in a Japanese population," *Forensic Sci Int*, vol. 172, no. 1, pp. 72–78, Oct. 2007, doi: 10.1016/J.FORSCIINT.2006.07.022.

[64] M. L. Pontes, L. Cainé, D. Abrantes, G. Lima, and M. F. Pinheiro, "Allele frequencies and population data for 17 Y-STR loci (AmpFlSTR Y-filer) in a Northern Portuguese population sample," *Forensic Sci Int*, vol. 170, no. 1, pp. 62–67, Jul. 2007, doi: 10.1016/J.FORSCIINT.2006.04.008.

[65] M. sen Shi, J. P. Tang, R. F. Bai, X. J. Yu, J. Y. Lv, and B. Hu, "Haplotypes of 20 Y-chromosomal STRs in a population sample from southeast China (Chaoshan area)," *Int J Legal Med*, vol. 121, no. 6, pp. 455–462, Nov. 2007, doi: 10.1007/S00414-007-0186-2.

[66] P. M. Domingues, L. Gusmão, D. A. da Silva, A. Amorim, R. W. Pereira, and E. F. de Carvalho, "Sub-Saharan Africa descendents in Rio de Janeiro (Brazil): population and mutational data for 12 Y-STR loci," *Int J Legal Med*, vol. 121, no. 3, pp. 238–241, May 2007, doi: 10.1007/S00414-007-0154-X.

[67] A. E. Decker, M. C. Kline, J. W. Redman, T. M. Reid, and J. M. Butler, "Analysis of mutations in father-son pairs with 17 Y-STR loci," *Forensic Sci Int Genet*, vol. 2, no. 3, Jun. 2008, doi: 10.1016/J.FSIGEN.2007.08.016.

[68] P. Sánchez-Diz *et al.*, "Population and segregation data on 17 Y-STRs: results of a GEP-ISFG collaborative study," *Int J Legal Med*, vol. 122, no. 6, pp. 529–533, Nov. 2008, doi: 10.1007/S00414-008-0265-Z.

[69] J. R. Padilla-Gutiérrez *et al.*, "Population data and mutation rate of nine Y-STRs in a mestizo Mexican population from Guadalajara, Jalisco, México," *Leg Med (Tokyo)*, vol. 10, no. 6, pp. 319–320, Nov. 2008, doi: 10.1016/J.LEGALMED.2008.03.004.

[70] M. Goedbloed *et al.*, "Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFlSTR Yfiler PCR amplification kit," *Int J Legal Med*, vol. 123, no. 6, pp. 471–482, Nov. 2009, doi: 10.1007/S00414-009-0342-Y.

[71] J. Ge, B. Budowle, X. G. Aranda, J. v. Planz, A. J. Eisenberg, and R. Chakraborty, "Mutation rates at Y chromosome short tandem repeats in Texas populations," *Forensic*

Sci Int Genet, vol. 3, no. 3, pp. 179–184, Jun. 2009, doi: 10.1016/J.FSIGEN.2009.01.007.

[72] S. H. Kim *et al.*, "Population genetics and mutational events at 6 Y-STRs in Korean population," *Forensic Sci Int Genet*, vol. 3, no. 2, Mar. 2009, doi: 10.1016/J.FSIGEN.2008.06.004.

[73] K. N. Ballantyne *et al.*, "Toward male individualization with rapidly mutating y-chromosomal short tandem repeats," *Hum Mutat*, vol. 35, no. 8, pp. 1021–1032, 2014, doi: 10.1002/HUMU.22599.

[74] C. Robino *et al.*, "Development of an Italian RM Y-STR haplotype database: Results of the 2013 GEFI collaborative exercise," *Forensic Sci Int Genet*, vol. 15, pp. 56–63, 2015, doi: 10.1016/J.FSIGEN.2014.10.008.

[75] Y. Wang *et al.*, "Genetic polymorphisms and mutation rates of 27 Y-chromosomal STRs in a Han population from Guangdong Province, Southern China," *Forensic Sci Int Genet*, vol. 21, pp. 5–9, Mar. 2016, doi: 10.1016/J.FSIGEN.2015.09.013.

[76] A. Adnan, A. Rakha, O. Lao, and M. Kayser, "Mutation analysis at 17 Y-STR loci (Yfiler) in father-son pairs of male pedigrees from Pakistan," *Forensic Sci Int Genet*, vol. 36, pp. e17–e18, Sep. 2018, doi: 10.1016/J.FSIGEN.2018.07.001.

[77] F. C. Bugoye, E. Mulima, and G. Misinzo, "Analysis of Mutation Rate of 17 Y-Chromosome Short Tandem Repeats Loci Using Tanzanian Father-Son Paired Samples," *Genet Res Int*, vol. 2018, 2018, doi: 10.1155/2018/8090469.

[78] E. Mertoglu, G. Filoglu, T. Zorlu, and O. Bulbul, "Estimation of the Y-chromosomal short tandem repeat (Y-STR) mutation rates in Turkey," *Turkish Journal of Biochemistry*, vol. 43, no. 2, pp. 142–150, Apr. 2018, doi: 10.1515/TJB-2017-0240/PDF.

[79] Y. Yang *et al.*, "Haplotypic polymorphisms and mutation rate estimates of 22 Y-chromosome STRs in the Northern Chinese Han father–son pairs," *Scientific Reports 2018 8:1*, vol. 8, no. 1, pp. 1–6, May 2018, doi: 10.1038/s41598-018-25362-3.

[80] W. Wu *et al.*, "Mutation rates at 42 Y chromosomal short tandem repeats in Chinese Han population in Eastern China," *Int J Legal Med*, vol. 132, no. 5, pp. 1317–1319, Sep. 2018, doi: 10.1007/S00414-018-1784-X.

[81] V. Petrovic, M. Kecmanovic, M. Keckarevic Markovic, and D. Keckarevic, "Assessment of mutation rates for PPY23 Y chromosome STR loci in Serbian father-son pairs,"

Forensic Sci Int Genet, vol. 39, pp. e5–e9, Mar. 2019, doi: 10.1016/J.FSIGEN.2018.11.014.

[82] J. Zhang *et al.*, "Mutation rates in father-son pairs of the 27 Y-STR loci in the Dezhou Han population from Shandong province, eastern China," *J Forensic Leg Med*, vol. 67, pp. 61–63, Oct. 2019, doi: 10.1016/J.JFLM.2019.06.015.

[83] M. Ay, A. Serin, H. Sevay, C. Gurkan, and H. Canan, "Genetic characterisation of 13 rapidly mutating Y-STR loci in 100 father and son pairs from South and East Turkey," *https://doi.org/10.1080/03014460.2018.1559353*, vol. 45, no. 6–8, pp. 506–515, Nov. 2019, doi: 10.1080/03014460.2018.1559353.

[84] L. Yuan *et al.*, "Mutation analysis of 13 RM Y-STR loci in Han population from Beijing of China," *Int J Legal Med*, vol. 133, no. 1, pp. 59–63, Jan. 2019, doi: 10.1007/S00414-018-1949-7.

[85] I. B. Ambrosio *et al.*, "Mutational data and population profiling of 23 Y-STRs in three Brazilian populations," *Forensic Sci Int Genet*, vol. 48, Sep. 2020, doi: 10.1016/J.FSIGEN.2020.102348.

[86] H. Lin, Q. Ye, P. Tang, T. Mo, X. Yu, and J. Tang, "Analyzing genetic polymorphism and mutation of 44 Y-STRs in a Chinese Han population of Southern China," *Leg Med (Tokyo)*, vol. 42, Feb. 2020, doi: 10.1016/J.LEGALMED.2019.101643.

[87] J. Fu, J. Cheng, C. Wei, M. A. Khan, Z. Jin, and J. Fu, "Assessing 23 Y-STR loci mutation rates in Chinese Han father-son pairs from southwestern China," *Mol Biol Rep*, vol. 47, no. 10, pp. 7755–7760, Oct. 2020, doi: 10.1007/S11033-020-05851-3.

[88] S. Bredemeyer, L. Roewer, and S. Willuweit, "Next generation sequencing of Y-STRs in father-son pairs and comparison with traditional capillary electrophoresis," *Forensic Sci Res*, 2021, doi: 10.1080/20961790.2021.1898078/SUPPL_FILE/TFSR_A_1898078_SM7293.XLSX.

[89] C. Vieira-Silva, P. Dario, T. Ribeiro, I. Lucas, H. Geada, and R. Espinheira, "Y-STR mutational rates determination in South Portugal Caucasian population," *Forensic Sci Int Genet Suppl Ser*, vol. 2, no. 1, pp. 60–61, Dec. 2009, doi: 10.1016/J.FSIGSS.2009.08.159.

[90] J. A. Soares-Vieira, A. E. C. Billerbeck, E. S. M. Iwamura, B. B. Mendonca, L. Gusmão, and P. A. Otto, "Population and mutation analysis of Y-STR loci in a sample from the city of São Paulo (Brazil)," *Genet Mol Biol*, vol. 31, no. 3, pp. 651–656, 2008, doi: 10.1590/S1415-47572008000400009.

[91] A. Laouina *et al.*, "Mutation rate at 17 Y-STR loci in 'Father/Son' pairs from moroccan population," *Leg Med*, vol. 15, no. 5, pp. 269–271, Sep. 2013, doi: 10.1016/J.LEGALMED.2013.03.003.

[92] B. M. Dupuy *et al.*, "Y-chromosome variation in a Norwegian population sample," *Forensic Sci Int*, vol. 117, no. 3, pp. 163–173, Apr. 2001, doi: 10.1016/S0379-0738(00)00397-2.

[93] W. Bar *et al.*, "DNA recommendations. Further report of the DNA Commission of the ISFH regarding the use of short tandem repeat systems. International Society for Forensic Haemogenetics," *Int J Legal Med*, vol. 110, no. 4, pp. 175–176, 1997, doi: 10.1007/S004140050061.

[94] "GRCh38.p14 - hg38 - Genome - Assembly - NCBI." https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.40 (accessed Apr. 25, 2022).

[95] K. N. Ballantyne *et al.*, "Mutability of Y-Chromosomal Microsatellites: Rates, Characteristics, Molecular Bases, and Forensic Implications," *Am J Hum Genet*, vol. 87, no. 3, p. 341, Sep. 2010, doi: 10.1016/J.AJHG.2010.08.006.

[96] T. J. Pemberton, C. I. Sandefur, M. Jakobsson, and N. A. Rosenberg, "Sequence determinants of human microsatellite variability," *BMC Genomics*, vol. 10, no. 1, pp. 1–19, Dec. 2009, doi: 10.1186/1471-2164-10-612/TABLES/6.

[97] M. T. Webster, N. G. C. Smith, and H. Ellegren, "Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments," *Proc Natl Acad Sci U S A*, vol. 99, no. 13, pp. 8748–8753, Jun. 2002, doi: 10.1073/PNAS.122067599/SUPPL_FILE/0675TABLE6.XLS.

[98] S. Martins, F. Calafell, V. C. Wong, J. Sequeiros, and A. Nio Amorim, "A multistep mutation mechanism drives the evolution of the CAG repeat at MJD/SCA3 locus," *European Journal of Human Genetics*, vol. 14, pp. 932–940, 2006, doi: 10.1038/sj.ejhg.5201643.

Appendix E

Antão-Sousa, S., Conde-Sousa, E., Gusmão, L., Amorim, A., & Pinto, N.

X-chromosomal STR mutation rate estimation biases and an informatics correction

Appendix 1: *p* values found in pairwise comparisons between the frequency of hidden mutations found when one single-, two- and three-steps mutation is simulated in parent(s)-child duos (mother-son: MS, mother-daughter: MD, father-daughter: FD) and trios (paternal and maternal mutation) for each marker and population.

For Appendix 1 please access document "Appendix E_X-chromosomal STR mutation rate estimation biases and an informatics correction" at:
https://drive.google.com/drive/folders/10uaPx_Bxmh2nJxrYxoSMMd7unp7Q0Mit?usp=share_link

Table S1: Markers showing the highest and lowest proportions (between parentheses) of hidden mutations for each population, and family type. One single-step mutation simulated in a parental meiosis of 1,000,000 family clusters of each, considering the allele frequencies of 12 X-STRs in three populations (Argentina, Eastern Asia, and Northern Europe) [35].

Population	Maker showing the highest proportion of hidden mutations				
	Duos			Trios	
	Mother-daughter	Father-daughter	Mother-son	Mutation in maternal meiosis	Mutation in paternal meiosis
Argentina	DXS8378 (0.613)	DXS8378 (0.326)	DXS8378 (0.214)	DXS8378 (0.214)	DXS8378 (0.069)
Eastern Asia	DXS7423 (0.692)	DXS7423 (0.435)	DXS7423 (0.237)	DXS7423 (0.237)	DXS7423 (0.099)
Northern Europe	DXS8378 (0.591)	DXS8378 (0.301)	DXS7423 (0.224)	DXS7423 (0.224)	DXS7423 (0.068)
Population	Maker showing the lowest proportion of hidden mutations				
	Duos			Trios	
	Mother-daughter	Father-daughter	Mother-son	Mutation in maternal meiosis	Mutation in paternal meiosis
Argentina	DXS10135 (0.178)	DXS10135 (0.065)	DXS10135 (0.062)	DXS10135 (0.062)	DXS10135 (0.005)
Eastern Asia	DXS10135 (0.197)	DXS10135 (0.072)	DXS10135 (0.071)	DXS10135 (0.070)	DXS10135 (0.006)
Northern Europe	DXS10135 (0.170)	DXS10135 (0.062)	DXS10135 (0.059)	DXS10135 (0.060)	DXS10135 (0.004)

Table S2: Average standard deviation and population showing the highest and lowest standard deviation values (in parentheses) regarding the proportion of hidden mutations across the different X-chromosomal markers for each family. One single-step mutation simulated in a parental meiosis of 1,000,000 family clusters of each, considering the allele frequencies of 12 X-STRs in three populations (Argentina, Eastern Asia, and Northern Europe) [35].

	Population showing the highest and the lowest standard deviation regarding proportion of hidden mutations				
Familial configuration	**Duos**			**Trios**	
	Mother-daughter	**Father-daughter**	**Mother-son**	**Mutation in maternal meiosis**	**Mutation in paternal meiosis**
Average	σ = 0.1480	σ = 0.0947	σ = 0.0557	σ = 0.0557	σ = 0.0250
Highest	Eastern Asia (σ = 0.1643)	Eastern Asia (σ = 0.1133)	Northern Europe (σ = 0.0576)	Northern Europe (σ = 0.0576)	Eastern Asia (σ = 0.0295)
Lowest	Argentina (σ = 0.1382)	Northern Europe (σ = 0.0831)	Argentina (σ = 0.0522)	Argentina (σ = 0.0522)	Argentina (σ = 0.0219)

Table S3: Populations showing the highest and the lowest average proportions (between parentheses) of hidden mutations, for each family type. One single-step mutation simulated in a parental meiosis of 1,000,000 family clusters of each, considering the allele frequencies of 12 X-STRs in three populations (Argentina, Eastern Asia, and Northern Europe) [35].

Population showing the highest average proportion of hidden mutations				
Duos			**Trios**	
Mother-daughter	**Father-daughter**	**Mother-son**	**Mutation simulated in maternal meiosis**	**Mutation simulated in paternal meiosis**
Eastern Asia (σ = 0.434)	Eastern Asia (σ = 0.214)	Eastern Asia (σ = 0.154)	Eastern Asia (σ = 0.154)	Eastern Asia (σ = 0.039)
Population showing the lowest average proportion of hidden mutations				
Duos			**Trios**	
Mother-daughter	**Father-daughter**	**Mother-son**	**Mutation simulated in maternal meiosis**	**Mutation simulated in paternal meiosis**
Argentina (σ = 0.380)	Argentina (σ = 0.174)	Argentina (σ = 0.131)	Argentina (σ = 0.131)	Argentina (σ = 0.028)

Table S4: Average standard deviation and marker showing the highest and lowest values (in parentheses) regarding the proportion of hidden mutations across the different X-chromosome markers for each family.

	Marker showing the highest and the lowest standard deviation regarding the proportion of hidden mutations				
	Duos			**Trios**	
	Mother-daughter	**Father-daughter**	**Mother-son**	**Maternal mutation**	**Paternal mutation**
Average	σ = 0.039	σ = 0.029	σ = 0.013	σ = 0.016	σ = 0.007
Highest	DXS10074 (σ = 0.098)	DXS7423 (σ = 0.087)	DXS10074 (σ = 0.044)	DXS10074 (σ = 0.044)	DXS7423 (σ = 0.021)
Lowest	DXS10101 (σ = 0.003)	DXS10101 (σ = 0.001)	DXS10101 (σ = 0.002)	DXS10101 (σ = 0.002)	DXS10101 (σ = 0.001)

* One single-step mutation simulated in a parental meiosis of 1,000,000 family clusters of each, considering the allele frequencies of 12 X-STRs in three populations (Argentina, Eastern Asia, and Northern Europe) [35].

Table S5: The average proportion of hidden mutations when two- and three-step mutations were simulated in mother-daughter, mother-son, father-daughter, and trios with daughter (mutation simulated in paternal and maternal meiosis), is presented.

Mutation simulated*	**Average proportion of hidden mutations**				
	Duos			**Trios (mutation simulated in)**	
	Mother-Daughter	**Mother-Son**	**Father-Daughter**	**Father**	**Mother**
2 steps	0.385	0.089	0.191	0.016	0.089
3 steps	0.363	0.049	0.191	0.007	0.049

* Two- and three-step single mutation simulated in a parental meiosis of 1,000,000 parent(s)-child duos and trios, considering the allele frequencies of 12 X-STRs in the populations of Argentina, Eastern Asia, and Northern Europe [35]. N=36,000,000, for each mutation step and family type.

Table S6: Markers showing the highest and lowest proportions (between parentheses) of hidden mutations for each population, mutation, and family type.

Population	Maker showing the highest proportion of hidden mutations									
	Two-step mutation*					Three-step mutation*				
	Duos			Trios		Duos			Trios	
	MD**	FD**	MS**	M**	F**	MD**	FD**	MS**	M**	F**
Argentina	DXS8378 (0.579)	DXS8378 (0.326)	DXS7132 (0.112)	DXS7132 (0.112)	DXS8378 (0.033)	DXS8378 (0.545)	DXS8378 (0.326)	DXS10103 (0.107)	DXS10103 (0.107)	DXS10103 (0.031)
E. Asia	DXS7423 (0.674)	DXS7423 (0.435)	DXS10079 (0.118)	DXS10079 (0.119)	DXS8378 (0.027)	DXS7423 (0.656)	DXS7423 (0.435)	DXS10103 (0.098)	DXS10103 (0.099)	DXS10103 (0.026)
N. Europe	DXS8378 (0.554)	DXS8378 (0.301)	DXS8378 (0.111)	DXS7132 (0.113)	DXS8378 (0.032)	DXS8378 (0.518)	DXS8378 (0.301)	DXS10103 (0.075)	DXS10103 (0.075)	DXS10103 (0.017)
Population	**Maker showing the lowest proportion of hidden mutations**									
	Two-step mutation*					**Three-step mutation***				
	Duos			**Trios**		**Duos**			**Trios**	
	MD**	FD**	MS**	M**	F**	MD**	FD**	MS**	M**	F**
Argentina	DXS10135 (0.175)	DXS10135 (0.065)	DXS10148 (0.560)	DXS10148 (0.553)	DXS10135 (0.004)	DXS10135 (0.172)	DXS10135 (0.065)	DXS8378 (0.013)	DXS8378 (0.013)	DXS8378 (0.002)
E. Asia	DXS10135 (0.192)	DXS10135 (0.072)	DXS7423 (0.039)	DXS7423 (0.038)	DXS10135 (0.005)	DXS10135 (0.187)	DXS10135 (0.072)	DXS7423 (0.006)	DXS7423 (0.006)	DXS7423 (0.001)
N. Europe	DXS10135 (0.168)	DXS10135 (0.062)	DXS10135 (0.057)	DXS10135 (0.057)	DXS10135 (0.004)	DXS10135 (0.165)	DXS10135 (0.062)	DXS8378 (0.017)	DXS8378 (0.017)	DXS8378 (0.003)

* Mutation simulated in a parental meiosis of 1,000,000 family configurations of each type and mutational steps, considering the allele frequencies of 12 X-STRs in three populations: Argentina, Eastern Asia, and Northern Europe [35].

**MD = mother-daughter duos, FD = father-daughter duos, MS = mother-son duos, M = trios with mutation simulated in the maternal meiosis, F = trios with mutation simulated in the paternal meiosis.

Table S7: Average standard deviation and population showing the highest and lowest values (in parentheses) regarding the proportion of hidden mutations across the different X-chromosome markers for each family and mutation type.

	Population showing the highest and the lowest standard deviation regarding proportion of hidden mutations									
	Two-step mutation*					**Three-step mutation***				
	Duos			**Trios**		**Duos**			**Trios**	
	MD**	FD**	MS**	M**	F**	MD**	FD**	MS**	M**	F**
Average	σ = 0.1405	σ = 0.0947	σ = 0.0227	σ = 0.0227	σ = 0.0087	σ = 0.1337	σ = 0.0947	σ = 0.0220	σ = 0.0221	σ = 0.0060
Highest	E. Asia (σ=0.1590)	E. Asia (σ=0.1173)	E. Asia (σ=0.0263)	E. Asia (σ=0.0263)	Argentina (σ=0.0093)	E. Asia (σ=0.1544)	E. Asia (σ=0.1173)	E. Asia (σ=0.0261)	E. Asia (σ=0.0263)	Argentina (σ=0.0075)
Lowest	Argentina (σ=0.1304)	N. Europe (σ=0.0831)	N. Europe (σ=0.0200)	N. Europe (σ=0.0201)	E. Asia (σ=0.0081)	Argentina (σ=0.1230)	N. Europe (σ=0.0831)	N. Europe (σ=0.0163)	N. Europe (σ=0.0164)	N. Europe (σ=0.0039)

*Mutation simulated in a parental meiosis of 1,000,000 family clusters of each type and mutational steps, considering the allele frequencies of 12 X-STRs in three populations (Argentina, Eastern Asia, and Northern Europe) [35].

**MD = mother-daughter duos, FD = father-daughter duos, MS = mother-son duos, M = trios with mutation simulated in the maternal meiosis, F = trios with mutation simulated in the paternal meiosis.

Table S8: Populations showing the highest and the lowest average proportions (between parentheses) of hidden mutations, for each mutation and family type.

Population showing the highest average proportion of hidden mutations									
Two-step mutation*					**Three-step mutation***				
Duos			**Trios**		**Duos**			**Trios**	
MD**	FD**	MS**	M**	F**	MD**	FD**	MS**	M**	F**
E. Asia (0.415)	E. Asia (0.214)	N. Europe (0.089)	N. Europe (0.089)	N. Europe (0.017)	E. Asia (0.393)	E. Asia (0.214)	Argentina (0.053)	Argentina (0.053)	Argentina (0.008)
Population showing the lowest average proportion of hidden mutations									
Two-step mutation*					**Three-step mutation***				
Duos			**Trios**		**Duos**			**Trios**	
MD**	FD**	MS**	M**	F**	MD**	FD**	MS**	M**	F**
Argentina (0.361)	Argentina (0.174)	Argentina (0.088)	Argentina (0.088)	Argentina (0.016)	Argentina (0.341)	Argentina (0.174)	N. Europe (0.045)	N. Europe (0.045)	N. Europe (0.006)

*Mutation simulated in a parental meiosis of 1,000,000 family configurations of each type and mutational steps, considering the allele frequencies of 10 autosomal STRs in three populations (Argentina, Eastern Asia, and Northern Europe) [35].

**MD = mother-daughter duos, FD = father-daughter duos, MS = mother-son duos, M = trios with mutation simulated in the maternal meiosis, F = trios with mutation simulated in the paternal meiosis

Table S9: Average standard deviation and marker showing the highest and lowest values (in parentheses) regarding the proportion of hidden mutations across the different X-chromosome markers for each family and mutation type.

	Marker showing the highest and the lowest standard deviation regarding the proportion of hidden mutations									
	Two-step mutation*					**Three-step mutation***				
	Duos			**Trios**		**Duos**			**Trios**	
	MD**	FD**	MS**	M**	F**	MD**	FD**	MS**	M**	F**
Average	σ=0.039	σ=0.029	σ=0.011	σ=0.011	σ=0.003	σ=0.038	σ=0.029	σ=0.008	σ=0.008	σ=0.002
Highest	DXS10074 (σ=0.093)	DXS7423 (σ=0.087)	DXS7423 (σ=0.038)	DXS7423 (σ=0.038)	DXS7423 (σ=0.009)	DXS7423 (σ=0.092)	DXS7423 (σ=0.087)	DXS7423 (σ=0.018)	DXS7423 (σ=0.018)	DXS10103 (σ=0.007)
Lowest	DXS10101 (σ=0.004)	DXS10101 (σ=0.001)	DXS7132 (σ=0.001)	DXS7132 (σ=0.001)	HPRTB (σ=0.0002)	DXS10101 (σ=0.004)	DXS10101 (σ=0.001)	DXS8378 (σ=0.002)	DXS8378 (σ=0.002)	DXS10135 (σ=0.0003)

*Simulated in a parental meiosis of 1,000,000 family clusters of each type and mutational steps, considering the allele frequencies of 12 X-STRs in three populations (Argentina, Eastern Asia, and Northern Europe) [35].

**MD = mother-daughter duos, FD = father-daughter duos, MS = mother-son duos, M = trios with mutation simulated in the maternal meiosis, F = trios with mutation simulated in the paternal meiosis.

2.2.3. Underestimation and misclassification of mutations at X chromosome STRs depend on population's allelic profile

Antão-Sousa, S., Conde-Sousa, E., Gusmão, L., Amorim, A., & Pinto, N. (2019). Underestimation and misclassification of mutations at X chromosome STRs depend on population's allelic profile. *Forensic Science International: Genetics Supplement Series*, *7*(1), 718-720. (https://doi.org/10.1016/j.fsigss.2019.10.150)

Forensic Science International: Genetics Supplement Series 7 (2019) 718–720

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

journal homepage: www.elsevier.com/locate/fsigss

Underestimation and misclassification of mutations at X chromosome STRs depend on population's allelic profile

S. Antão-Sousa[a,b,c,d,*], E. Conde-Sousa[c,e], L. Gusmão[d], A. Amorim[a,b,c], N. Pinto[b,c,f]

[a] Faculty of Sciences of the University of Porto (FCUP), Portugal
[b] Institute of Pathology and Molecular Immunology from University of Porto (IPATIMUP), Portugal
[c] Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, Portugal
[d] DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Brazil
[e] Instituto Nacional de Engenharia Biomédica (INEB), Porto, Portugal
[f] Center of Mathematics of the University of Porto (CMUP), Portugal

ARTICLE INFO

Keywords:
X chromosome
STRs
Mutation rates
Estimates
Mutation mechanism

ABSTRACT

Short tandem repeats are extremely abundant in the human genome and widely used in population and forensic genetics. The conventional approach to estimate mutation rates for STRs involves the detection of Mendelian incompatibilities and their frequency quantification, in family duos or trios. In this work, we aim to measure the impact this estimation methodology implies in the underestimation of mutation rates.

We computed two independent experiences simulating one single-step mutation and one double step mutation in each pedigree. The pedigree with the highest number of compatibilities observed (hidden mutations) was mother-daughter duo, followed by the father-daughter duo, mother-son duo and, finally, mother-father-child trio, for the one and two-steps simulations. The estimation of X-STRs mutation rates from simple counting proved to be biased through systematic underestimation in an extent which depends on allele frequencies distribution of the marker, type of pedigree and number of repeat units involved.

1. Introduction

Short tandem repeats (STRs) have been widely used in X chromosome population and forensic genetics. Their polymorphism being generated by mutation, the study of mutational mechanisms and rates is essential for their sensible use.

The primary mutational mechanism leading to changes in STRs length is thought to be polymerase template slippage [1,2]. Single step mutations, gains or losses of one repeat, are assumed to be preponderant and the likelihood of a mutation is thought to be inversely correlated with the number of repeats differing between parental and filial alleles. Traditionally, average mutation rates are estimated per marker through the computation of the proportion of Mendelian incompatibilities observed in trios (father, mother and child) or duos (parent and child). Nevertheless, the genotypic observations may be explained by alternative events. Specifically, a mutation may not result in a Mendelian incompatibility – the so called "hidden mutations", or a genotypic configuration can be often explained through different mutations involving various numbers of repeat units, the one involving the smallest number being traditionally considered. This naïve approach leads to the underestimation of multistep mutations and of average mutation rates in general, and several studies [3–6] have tackled these issues for autosomal STRs, presenting also some approaches to reduce the bias.

In this work, and resorting to computational simulations, we assess the impact of the estimation of mutation rates through the proportion of Mendelian compatibilities in duos and trios at X-STRs.

2. Material and methods

In this work we considered 8 X-STRs: DXS10134, DXS8378, DXS7132, DXS10101, DXS10074, DXS10135, HPRTB and DXS7423, and a Norwegian population database [7]. Resorting to Python programming language, we created 1,000,000 joint genotypic configurations, for each marker and pedigree: father-daughter, mother-daughter and mother-son duos, and father-mother-daughter trios.

For each of the 32,000,000 joint genotypic configurations simulated, we computed two independent scenarios simulating: (i) one single-step mutation, and (ii) one double step mutation, randomly assigned to one of the parents' alleles in case of trios. Also, the gain or loss

* Corresponding author at: Faculty of Sciences of the University of Porto (FCUP), Portugal.
E-mail address: sofias@ipatimup.pt (S. Antão-Sousa).

https://doi.org/10.1016/j.fsigss.2019.10.150
Received 6 September 2019; Received in revised form 10 October 2019; Accepted 11 October 2019
Available online 21 October 2019
1875-1768/

Table 1
Frequency of compatibilities ("hidden" mutations) observed (Comp – Mendelian compatibility), frequency of one-step mutations when double-step mutations were simulated (1-step IM – one-step identified mutation), frequency of actual mutation identification (2-step IM – two-steps identified mutation), for each pedigree considered and for each marker and heterozygosity per marker.

	Markers	Father-daughter duos		Mother-daughter duos		Mother-son duos		Mother-father-daughter trios		Heterozygosity
		Comp	1-step IM	Comp	1-step IM	Comp	1-step IM	Comp	1-step IM	
One-step mutation simulated	DXS10134	0.57	0.43	0.67	0.33	0.13	0.87	0.08	0.92	0.85
	DXS8378	0.65	0.35	0.79	0.21	0.22	0.78	0.14	0.86	0.70
	DXS7132	0.63	0.37	0.77	0.23	0.21	0.79	0.13	0.87	0.74
	DXS10101	0.55	0.45	0.63	0.37	0.09	0.91	0.05	0.95	0.90
	DXS10074	0.58	0.42	0.69	0.31	0.12	0.88	0.07	0.93	0.84
	DXS10135	0.54	0.46	0.60	0.40	0.07	0.93	0.04	0.96	0.93
	HPRTB	0.63	0.37	0.77	0.23	0.21	0.79	0.13	0.87	0.74
	DXS7423	0.64	0.36	0.78	0.22	0.22	0.78	0.14	0.86	0.72
	Average	**0.60**	**0.40**	**0.71**	**0.29**	**0.16**	**0.84**	**0.10**	**0.90**	
	Standard Deviation	**0.04**		**0.07**		**0.06**		**0.04**		

	Markers	Father-daughter duos			Mother-daughter duos			Mother-son duos			Mother-father-daughter trios			Heterozygosity
		Comp	1-step IM	2-step IM	Comp	1-step IM	2-step IM	Comp	1-step IM	2-step IM	Comp	1-step IM	2-step IM	
Two-step mutations simulated	DXS10134	0.57	0.13	0.30	0.66	0.18	0.15	0.10	0.19	0.72	0.05	0.12	0.82	0.85
	DXS8378	0.65	0.22	0.14	0.77	0.18	0.04	0.11	0.24	0.65	0.07	0.17	0.76	0.70
	DXS7132	0.63	0.21	0.16	0.75	0.20	0.05	0.11	0.25	0.64	0.07	0.18	0.75	0.74
	DXS10101	0.55	0.09	0.36	0.62	0.17	0.21	0.06	0.13	0.81	0.04	0.08	0.89	0.90
	DXS10074	0.58	0.12	0.30	0.67	0.18	0.15	0.07	0.15	0.78	0.04	0.10	0.86	0.84
	DXS10135	0.54	0.07	0.40	0.59	0.15	0.26	0.06	0.13	0.81	0.04	0.07	0.89	0.93
	HPRTB	0.63	0.21	0.16	0.75	0.19	0.06	0.10	0.25	0.65	0.06	0.17	0.76	0.74
	DXS7423	0.64	0.22	0.14	0.76	0.19	0.04	0.11	0.25	0.64	0.07	0.18	0.75	0.72
	Average	**0.60**	**0.16**	**0.24**	**0.70**	**0.18**	**0.12**	**0.09**	**0.20**	**0.71**	**0.05**	**0.14**	**0.81**	
	Standard deviation	**0.04**	**0.06**	**0.11**	**0.07**	**0.02**	**0.08**	**0.02**	**0.05**	**0.08**	**0.01**	**0.05**	**0.06**	

of repeats were randomly considered.

For both cases we computed the frequency of cases that resulted in Mendelian compatibilities ("hidden" mutations), and for (ii) we also weighted the proportion of genotypic configurations that, despite resulting from a double step mutation, were also explainable by a single step one. The standard deviation between the proportions obtained for the different markers was also computed.

3. Results

When single-step mutations were simulated, two possible outcomes for the resulting genotypic configurations were considered: the observation of a Mendelian compatibility or otherwise. In this case, because a one-step mutation in every allelic transmission was simulated under the assumption of codominance, the single step mutation would be correctly identified whenever an incompatibility was observed. Nevertheless, ambiguity on the gain or loss of the repeat may occur. The pedigree with the highest number of compatibilities observed was mother-daughter duo (71%), followed by the father-daughter duo (60%), mother-son duo (16%) and, finally, mother-father-daughter trio (10%).

On the other hand, when two-step mutations were simulated, we considered three possible outcomes for the resulting genotypic configuration: compatibility with (i) Mendelian transmission, (ii) a single step, or (iii) a double step mutation. In this case, because we simulated a two-step mutation in every allelic transmission of one parent, the mutation would be correctly identified only in the cases where a two-step mutation was the explanation involving the least number of repeats. Also, in this case, ambiguity on gain or loss is possible. As before, the pedigree with the highest number of compatibilities observed was the mother-daughter duo (70%), followed by the father-daughter duo (60%), the mother-son duo (9%) and, finally, the mother-father-daughter trio (5%).

The standard deviation is, for most pedigrees analyzed, higher when two-step mutations were simulated.

Globally, it is harder to detect two step mutations than single step ones, and the number of resulting compatible genotypic configurations is slightly lower when double step mutations were simulated.

In all cases, the markers that showed the highest level of mutation detection were DXS10135 and DXS10101, which are the two markers with the highest heterozygosity. Conversely, those showing the lowest detection capacity were DXS8378 and DXS7423, with the lowest heterozygosity. See Table 1.

4. Discussion

As expected, the capacity to detect mutations through the analysis of Mendelian incompatibilities was greater in trios than in duos. Comparing mother-son and father-daughter duos, the first had the best detection rate as the pool of possible parental alleles was reduced to two, contrarily to the second. Mother-daughter duos had the lowest detection rate as the ambiguity resides in assigning which parental allele originated which filial one (equivalent to the autosomal mode of transmission).

As expected, some mutations would not be identified through the analysis of Mendelian incompatibilities, and some two-step ones would be interpreted as single step. When two-step mutations were simulated, mutation detection rate increased in comparison to the one-step mutations simulations.

Some markers, like DXS10135 and DXS10101, the two markers with the highest heterozygosity, allowed for better detection of mutation than others, like DXS8378 and DXS7423, with the lowest heterozygosity, which had the lowest mutation detection rates.

5. Conclusion

Due to the haplodiploid X-chromosome mode of transmission mutation detection capacity is higher than for autosomes and there is less ambiguity in the assignment of parental and filial alleles. However, biases in the rate estimation and classification are still an issue. Our results demonstrate that for X chromosomal STRs (i) current Mendelian incompatibility, maximum parsimony method, underestimates both overall and multistep mutation rates, and (ii) show the effect of allele frequency distribution on these parameters, demonstrating that mutation detection depends on the allelic diversity at the locus in the population under study.

An important consequence is, therefore, that mutation rates as estimated by simple incompatibilities counting method provide different

values across populations varying significantly in the degree of polymorphism.

The authors declare no conflict of interests.

Acknowledgements

This work was partially financed by FEDER—Fundo Europeu de Desenvolvimento Regional - funds through the COMPETE 2020—Operacional Programme for Competitiveness and Internationalisation (POCI), Portugal 2020; Portuguese funds through FCT—Fundação para a Ciência e a Tecnologia/Ministério da Ciência, Tecnologia e Inovação in the framework of the Decreto-Lei nº 57/2016 de 29 de Agosto (NP work contract) and through the doctoral grant SFRH/BD/136284/2018; and the projects "Institute for Research and Innovation in Health Sciences" (POCI-01-0145-FEDER-007274) and "Center of Mathematics of the University of Porto" (UID/MAT/00144/2019); FCT National Roadmap of Research Infrastructures project "PPBI: PPBI - Portuguese Platform of BioImaging" PPBI-POCI-01-0145-FEDER-022122. LG (ref. 305330/2016-0) was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq and Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro – FAPERJ (CNE-2018).

References

[1] C. Schlötterer, D. Tautz, Slippage synthesis of simple sequence DNA, Nucleic Acids Res. 20 (2) (1992) 211–215.
[2] M. Strand, T.A. Prolla, R.M. Liskay, T.D. Petes, Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair, Nature 365 (6443) (1993) 274.
[3] R. Chakraborty, D.N. Stivers, Y. Zhong, Estimation of mutation rates from parentage exclusion data: applications to STR and VNTR loci, Mutat. Res. Mol. Mech. Mutagen. 354 (1) (1996) 41–48.
[4] P. Vicard, A.P. Dawid, A statistical treatment of biases affecting the estimation of mutation rates, Mutat. Res. Mol. Mech. Mutagen. 547 (1–2) (2004) 19–33.
[5] P. Vicard, A.P. Dawid, J. Mortera, S.L. Lauritzen, Estimating mutation rates from paternity casework, Forensic Sci. Int. Genet. 2 (1) (2008) 9–18.
[6] K. Slooten, F. Ricciardi, Estimation of mutation probabilities for autosomal STR markers, Forensic Sci. Int. Genet. 7 (3) (2013) 337–344.
[7] B.M. Dupuy, M. Stenersen, D. Kling, Frequency data for 35 autosomal STR markers in a Norwegian, an East African, an East Asian and Middle Asian population and simulation of adequate database size, Forensic Sci. Int. Genet. Suppl. Ser. 4 (1) (2013) e378–e379.

2.2.4. How often have X- and autosomal-STRs mutations equivocal parental origin been assigned?

Antão-Sousa, S., Conde-Sousa, E., Gusmão, L., Amorim, A., & Pinto, N. (2022). How often have X-and autosomal-STRs mutations equivocal parental origin been assigned?. *Forensic Science International: Genetics Supplement Series*, *8*, 99-101. (https://doi.org/10.1016/j.fsigss.2022.09.035)

Forensic Science International: Genetics Supplement Series 8 (2022) 99–101

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

journal homepage: www.elsevier.com/locate/fsigss

How often have X- and autosomal-STRs mutations equivocal parental origin been assigned?

Sofia Antão-Sousa [a,b,c,d], Eduardo Conde-Sousa [a,e], Leonor Gusmão [d], António Amorim [a,b,c], Nádia Pinto [a,b,f,*]

[a] Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, Porto, Portugal
[b] Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal
[c] Department of Biology, Faculty of Sciences of University of Porto (FCUP), Porto, Portugal
[d] DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Rio de Janeiro, Brazil
[e] Instituto de Engenharia Biomédica (INEB), Porto, Portugal
[f] Center of Mathematics of University of Porto (CMUP), Porto, Portugal

ARTICLE INFO

Keywords:
Mutation modeling
Paternal mutations
Maternal mutations
Microsatellites
X-STRs
Mutation

ABSTRACT

Short tandem repeat markers (STRs) are widely applied in population, evolutionary, and forensic genetics, due to extensive polymorphism in the number of repetitive motifs. The primary mutational mechanism leading to changes in the length of STRs is thought to be polymerase template slippage. Mutation rates in STRs and corresponding parental assignment are usually assessed through the number of Mendelian incompatibilities observed in one-generational, parent(s)-child, pedigrees, and paternal mutations have been assumed to be preponderant over maternal ones. Notwithstanding, diploid and haplodiploid modes of genetic transmission may not allow for the unequivocal assigning of the mutation to the correct parental origin (either paternal or maternal), especially when genotyping methodologies of fragment length determination are employed. In this work, the frequency under which a mutation might be assigned to the wrong parental origin or be interpreted as having an ambiguous origin is analyzed for both diploid and haplodiploid modes of genetic transmission. Genotypic configurations were generated with Python™ programming language, considering parents-child trios for autosomal transmission, and parents-daughter trios for the X chromosomal one. One single-, one two- or one three-step mutation was simulated in each familial constellation, and the resulting genotypic configuration was analyzed regarding the parental assignment of the mutation. When considering autosomal transmission, the meiosis suffering mutation was randomly selected. Contrarily, differential analyses were performed for paternal and maternal mutations for X-chromosomal transmission. In this work, we show that the biases in the rates between paternal and maternal mutations differ for autosomal and X-chromosomal modes of transmission. In the differential analysis performed for the X-chromosomal STRs, it is possible to ascertain that the maternal and paternal meioses are subject to different biases, the latter being better estimated than the first. This work shows that simulated data, along with reliable and properly communicated real one, may be crucial for the correct modeling of biological processes, such as the mutation in STRs.

1. Introduction

Short tandem repeats (STRs) or microsatellites are highly mutating repetitive stretches of DNA of usually 2–6 base pairs. The elevated polymorphism resulting from the high mutation rate of these genetic markers makes them extremely useful in several contexts, such as forensic, evolutionary, and population genetics [1–3]. The understanding of the mutational dynamics is, thus, crucial for the quantification of the data. The polymerase template slippage is thought to be the main mechanism behind length mutations where one or several repeats – single- and multistep mutations, respectively, are gained or lost in the derived sequence [4,5]. Mutation rates are usually estimated through the simple counting of Mendelian incompatibilities in one generation pedigrees, parents-child trios or parent-child duos, comparative to the number of genetic transmissions observed. This implies, for both the autosomal and X-chromosomal modes of transmission, the possibility of

* Corresponding author at: Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, Porto, Portugal.
E-mail address: npinto@ipatimup.pt (N. Pinto).

https://doi.org/10.1016/j.fsigss.2022.09.035
Received 19 September 2022; Accepted 29 September 2022
Available online 30 September 2022

Table 1
Rates of mutations that resulted in Mendelian incompatibilities, in which parental meiosis would be ambiguous or wrongly assigned, when either autosomal-STRs were analysed in parents-child trios or X-STR mutations were analysed in parents-daughter trios.

Autosomal mode of transmission										
Mutation change	**One-step**		**Two-steps**				**Three-steps**			
Parental origin observed	**Hidden**	**Ambiguous**	**Hidden**	**Ambiguous**		**Wrong parent (one-step)**	**Hidden**	**Ambiguous**		**Wrong parent (one- or two-steps)**
				Two-step	**One-step**			**Three-steps**	**One- or two-steps**	
Random mutation	16.5%	17.8%	11.1%	12.1%	4.7%	3.1%	7.3%	11.5%	5.1%	4.6%

X-chromosomal mode of transmission										
Mutation change	**One-step**		**Two-steps**				**Three-steps**			
Parental origin observed	**Hidden**	**Ambiguous**	**Hidden**	**Ambiguous**		**Wrong parent (one-step)**	**Hidden**	**Ambiguous**		**Wrong parent (one- or two-steps)**
				Two-step	**One-step**			**Three-steps**	**One- or two-steps**	
Maternal mutation	14.4%	17.4%	8.9%	13.6%	4.1%	–	4.9%	13.0%	5.4%	–
Paternal mutation	3.4%	15.7%	1.6%	13.3%	–	4.1%	0.7%	16.7%	–	8.2%

erroneous mutation assignment, either in the number of repeats or the parental origin, especially when genotyping methodologies of fragment length determination are used.

For all nuclear modes of genetic transmission (autosomal, X-chromosomal, and Y-chromosomal), it has been generally accepted for STRs that: i. single-step mutations are much more frequent than multistep ones [6]; ii. longer alleles are more mutable than shorter ones; iii. the paternal genetic transmission is more prone to mutation than the maternal one (with the obvious exception for the Y chromosome, with exclusive patrilineal inheritance) [7]; and iv. the paternal age is positively correlated with the occurrence of mutations [8]. To assess the strength of these assumptions, biological, reliable, and properly reported data are needed. However, simulated data is also critical for the understanding and modeling of these phenomena. In this work, we focus our attention on the assumption that paternal mutations are more common than the maternal ones, showing how the mode of transmission considered, or even the interpretation of the expert based on priors may bias this conclusion.

2. Material and methods

Genotypic configurations were generated with Python™ programming language, considering parents-child trios for autosomes, and parents-daughter trios for the X chromosome. One single-, one two-, or one three-step mutation was simulated in each familial constellation. For autosomal transmission, the meiosis suffering mutation was randomly selected, but when considering X-chromosomal transmission, differential analyses were performed for paternal and maternal mutations. For each case, 1,000,000 genotypic trios were simulated. The frequency under which a mutation might be assigned to the wrong parental origin or interpreted as having ambiguous parental origin was computed and analyzed.

3. Results and conclusions

When a single-step mutation was simulated in mother-father-child trios under the scope of autosomal transmission, 83.5% of mutations were detected (not hidden), and out of these 17.8% would be interpreted as having an ambiguous parental origin. When a two-step mutation was simulated, 88.9% of mutations were detected and 16.8% of these would be interpreted as having an ambiguous parental origin, and 3.1% would be interpreted as single-step paternal mutations (see Table 1).

When a single-step maternal mutation was simulated in mother-father-daughter trios assuming the X-chromosomal transmission, 85.6% of the mutations were detected, and 17.4% out of these would be interpreted as having ambiguous parental origin. In the case of a paternal mutation simulated, 96.6% were detected and 13.3% would be interpreted as having ambiguous parental origin (see Table 1).

The biases in the rates between paternal and maternal mutations differ for autosomal and X-chromosomal modes of transmission. For autosomal transmission, biases in the evaluation of paternal and maternal mutation rates may occur due to the expert's interpretation based on priors.

On the other hand, considering haplodiploid markers, biases associated with the different paternal and maternal modes of genetic transmission should also be considered as the level of detection of maternal and paternal mutations differs. Indeed, regardless of the interpretation of the expert, maternal hidden mutations showed to be 4.2 times more frequent than the paternal ones (in the case of single-step mutation, 5.6 for two-steps and 7 for three-steps). This can bring to the discussion the true ratio between paternal and maternal mutations. Indeed, since X chromosomal paternal mutations are more detectable than maternal ones, as shown in this work, the differential between the two rates might be overestimated.

Acknowledgments

This work was partially financed by FEDER-Fundo Europeu de Desenvolvimento Regional funds through the COMPETE 2020 – Operational Program for Competitiveness and Internationalization (POCI), Portugal 2020, and by Portuguese funds through FCT-Fundação para a Ciência e a Tecnologia/Ministério da Ciência, Tecnologia e Inovação in the framework of the projects "Institute for Research and Innovation in Health Sciences" (POCI-01-0145-FEDER-007274). NP is supported by FCT, under the program contract provided in Decree-Law no. 57/2016 of August 29. SAS is funded by a FCT doctoral grant SFRH/BD/136284/2018. LG is supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (ref. 306342/2019-7), and Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro – FAPERJ (CNE-2022).

Conflict of interests

The authors declare no conflict of interest.

References

[1] R. Flores-Espinoza, et al., Investigating genetic diversity in admixed populations from Ecuador, Am. J. Phys. Anthropol. 176 (1) (2021) 109–119, https://doi.org/10.1002/AJPA.24341.

[2] S. Srithawong, et al., Close genetic relationship between central Thai and Mon people in Thailand revealed by autosomal microsatellites, Int. J. Leg. Med. 135 (2) (2021) 445–448, https://doi.org/10.1007/S00414-020-02290-4.
[3] G. He, Z. Wang, M. Wang, Y. Hou, Genetic diversity and phylogenetic differentiation of Southwestern Chinese Han: a comprehensive and comparative analysis on 21 non-CODIS STRs, Sci. Rep. 7 (1) (2017), https://doi.org/10.1038/S41598-017-13190-W.
[4] M. Strand, T.A. Prolla, R.M. Liskay, T.D. Petes, Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair, *1993 365: 6443*, Nature 365 (6443) (1993) 274–276, https://doi.org/10.1038/365274a0.
[5] C. Schlötterer, D. Tautz, Slippage synthesis of simple sequence DNA, Nucleic Acids Res. 20 (2) (1992) 211, https://doi.org/10.1093/NAR/20.2.211.
[6] M. Wierdl, M. Dominska, T.D. Petes, Microsatellite instability in yeast: dependence on the length of the microsatellite, Genetics 146 (3) (1997) 769–779, https://doi.org/10.1093/GENETICS/146.3.769.
[7] N. Pinto, V. Pereira, C. Tomas, S. Loiola, E.F. Carvalho, N. Modesti, L. Gusmão, Paternal and maternal mutations in X-STRs: a GHEP-ISFG collaborative study, Forensic Sci. Int. Genet. 46 (2020), 102258.
[8] J.X. Sun, et al., A direct characterization of human mutation based on microsatellites, Nat. Genet. 44 (10) (2012) 1161–1165, https://doi.org/10.1038/NG.2398.

2.2.5. How frequently are autosomal and X-chromosomal STRs multistep mutations perceived as having fewer steps?

Antão-Sousa, S., Conde-Sousa, E., Gusmão, L., Amorim, A., & Pinto, N. (2022). How frequently are Autosomal and X-STRs multistep mutations perceived as single-step?. *Forensic Science International: Genetics Supplement Series*, *8*, 165-166. (https://doi.org/10.1016/j.fsigss.2022.10.022)

Forensic Science International: Genetics Supplement Series 8 (2022) 165–166

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

journal homepage: www.elsevier.com/locate/fsigss

How frequently are Autosomal and X-STRs multistep mutations perceived as single-step?

Sofia Antão-Sousa [a,b,c,d,*], Eduardo Conde-Sousa [a,e], Leonor Gusmão [d], António Amorim [a,b,c], Nádia Pinto [a,b,f]

[a] Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, Porto, Portugal
[b] Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal
[c] Department of Biology, Faculty of Sciences of University of Porto (FCUP), Porto, Portugal
[d] DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Rio de Janeiro, Brazil
[e] Instituto de Engenharia Biomédica (INEB), Porto, Portugal
[f] Center of Mathematics of University of Porto (CMUP), Porto, Portugal

ARTICLE INFO

Keywords:
Autosomal STRs
X-STRs
Mutation estimates biases
STRs mutation rates
Mutation modeling

ABSTRACT

Short tandem repeats (STRs) incur in length mutations that involve the loss or gain of repeats. STR mutation rates are usually estimated considering the rates of observed Mendelian incompatibilities in one generation familial configurations. When considering multistep mutations, for the autosomal and X-chromosomal modes of genetic transmission, underestimations are inevitable when using this approach (MIA), due to the occurrence of mutational events deceptively perceived as involving fewer steps. The rate of this occurrence depends on the mode of genetic transmission considered, the parental origin of the mutation, the type of familial configuration considered, and the genotypic background of the progenitor(s). MIA biases were weighted and compared for the diploid and haplodiploid modes of transmission, using familial genotypic configurations (parent(s)-child duos and trios) generated resorting to Python™ and real population databases from Norway, Somalia, and Spain for 10 Aut-STRs and Argentina, Eastern Asia, and Northern Europe for 12 X-STRs. One two- or one three-step mutation was simulated in each of the 1,000,000 familial configurations. The frequency with which mutations could be interpreted as involving fewer steps, when the most parsimonious reasoning is employed, was computed. Results showed that the magnitude and type of biases depend on the type of familial data and the genetic mode of transmission considered, being higher in duos than in trios, both in autosomes and the X chromosome. Indeed, whether X- or Aut-STRs are analyzed, trios generally provide better estimates and should be favored over duos. The pooling of the two types of data is not advised. The greater the number of steps involved in the mutation, the worst the estimates obtained. In X-chromosomal analyzes, trios with a paternal mutation presented the best estimates and mother-daughter duos the worst; mother-son duos showed similar estimates to trios when a maternal mutation was considered.

1. Introduction

Short tandem repeats (STRs) are highly repetitive DNA sequences of 2–6 base pairs, ubiquitous in the human genome, and one of the most used genetic markers in forensic genetics. With high mutation rates that result in extensive polymorphism, these markers are susceptible to length mutations involving the gain or loss of repeats. The mutation mechanism thought to be operating in length mutations is the slippage of the DNA polymerase. Mutation rate estimates, essential for the appraisal of genetic profiling results, rely on the simple counting of Mendelian incompatibilities observed in one-generation familial configurations - Mendelian incompatibility approach (MIA). The MIA rests on the assumption that mutations involving fewer steps are much more probable, hence, the scenario which encompasses the least changes (no mutation or mutation with the least number of steps possible) is the one considered. Therefore, through the MIA, for the autosomal and X-chromosomal markers, there is the possibility that the genotypic familial configuration does not allow the observation of the mutation that occurred or induces a wrong interpretation of the genotypic scenario and mutation occurred. This noticeably results in biased estimations.

* Corresponding author at: Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, Porto, Portugal.
E-mail address: sofias@ipatimup.pt (S. Antão-Sousa).

https://doi.org/10.1016/j.fsigss.2022.10.022
Received 19 September 2022; Accepted 17 October 2022
Available online 18 October 2022

Table 1
Rates of two- and three-step mutations that would be interpreted as involving fewer mutational steps (SM) and rates of hidden mutations (HM) when Aut- and X-STRs were analyzed.

Genetic mode of transmission												
Autosomal	Mutation steps simulated	Trios				Duos						
		SM		HM		SM		HM				
	Two	22.5%		11.1%		33.7%		38.8%				
	Three	30.4%		7.4%		47.1%		36.9%				
X-chromosomal	Mutation steps simulated	Trios (mutation simulated in)				Duos						
		Father		Mother		Mother-daughter		Mother- son		Father-daughter		
		SM	HM	SM	HM	SM	HM	SM	HM	SM	HM	
	Two	6.8%	1.6%	19.3%	8.9%	34.9%	38.5%	19.3%	8.9%	28.8%	19.1%	
	Three	10.7%	0.7%	21.5%	4.9%	49.2%	36.3%	26.9%	4.9%	46.6%	19.1%	

SM - proportion of mutations interpreted as involving fewer steps.
HM - proportion of hidden mutations observed.

This issue has been previously considered for autosomal STRs, with mathematical and informatics tools presented to correct per locus mutation rates estimation and computation of the misidentification of multistep mutations, allowing the correction of estimations obtained through the MIA [1,2].

2. Material and methods

In this work, we aimed to quantify the biases associated with the simple counting of observed incompatibilities to estimate mutation rates in autosomal and X-chromosomal STRs. Namely, how frequently multistep mutations are interpreted as involving fewer steps or none (hidden mutation).

Genealogical parent(s)-child genotypic configurations were generated using real population databases where a single mutation of either two or three steps was always simulated in each of the 1,000,000 meiosis simulated.

Parent(s)-child duos and trios configurations were simulated for 10 autosomal STRs using real allele frequencies from the populations of Norway, Somalia, and Spain [3].

Parents-daughter trios and mother-daughter, mother-son, and father-daughter duos familial configurations were simulated for 12 X-STRs using real allele frequencies from the populations of Argentina, Eastern Asia, and Northern Europe [4].

In this work, we present the rate of cases where multistep (two- and three-steps) mutations resulted either in Mendelian incompatibilities that would be interpreted as involving fewer mutational steps, or in Mendelian compatibilities (hidden mutations).

3. Results

For the case of the autosomal mode of transmission in duos, results show that as high as 47.1% of three-step mutations might be wrongly interpreted as involving fewer number of steps and 36.9% of three-step mutations were hidden, with no genotypic incompatibilities observed – see Table 1. For trios, the proportion of mutations interpreted as involving fewer number of steps drops to 30.4% and 7.4% for hidden mutations. Noticeably, trios provide much better estimates than duos.

For the case of the X-chromosomal mode of transmission, mother-daughter duos provided the worst estimates, with 49.2% of three-step mutations interpretable as involving fewer steps and 36.3% resulting in hidden mutations, as was expected since the mode of transmission is like that of the autosomes (two parental alleles and two filial alleles). The duo which presented the best estimates was the mother-son duo. The genotypic configuration that obtained the best estimates was trios with a paternal mutation simulated, where only 8.4% of two-step mutations would be wrongly interpreted (fewer steps or hidden mutation). Mother-son duos presented similar estimates to trios when a maternal mutation was simulated.

4. Discussion and conclusions

The accuracy of STR mutation rates estimated through the proportion of Mendelian incompatibilities observed in one-generation familial configurations varies widely considering the type of family data considered.

Multistep mutations are systematically underestimated when incompatibilities observed are interpreted as involving the gain or loss of the smallest possible number of repeats, as is the standard practice. The use of parent-child duos to estimate average mutation rates for diploid and haplodiploid modes of genetic transmission entails a high proportion of expected hidden mutations and difficulties in the correct mutation assignment, particularly in the case of parent-child for autosomes and mother-daughter duos for X-chromosomal markers.

In most cases, biases are more important when duos are considered. As such, pooling of data from duos and trios should not be performed.

Conflict of interest statement

The authors declare no conflict of interest.

Acknowledgments

This work was partially financed by FEDER- Fundo Europeu de Desenvolvimento Regional funds through the COMPETE 2020 - Operational Program for Competitiveness and Internationalization (POCI), Portugal 2020, and by Portuguese funds through FCT- Fundação para a Ciência e a Tecnologia/Ministério da Ciência, Tecnologia e Inovação in the framework of the projects "Institute for Research and Innovation in Health Sciences" (POCI-01-0145-FEDER-007274). NP is supported by FCT, under the program contract provided in Decree-Law no. 57/2016 of August 29. SAS is funded by a FCT doctoral grant SFRH/BD/136284/2018. LG is supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (ref. 306342/2019-7), and Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro - FAPERJ (CNE-2022).

References

[1] K. Slooten, F. Ricciardi, Estimation of mutation probabilities for autosomal STR markers, Forensic Sci. Int. Genet. 7 (2013) 337–344, https://doi.org/10.1016/J.FSIGEN.2013.01.006.
[2] S. Antão-Sousa, E. Conde-Sousa, L. Gusmão, A. Amorim, N. Pinto, Estimations of mutation rates depend on population allele frequency distribution: the case of autosomal microsatellites, Genes 13 (7) (2022) 1248.
[3] Daniel Kling, Andreas O. Tillmar, Thore Egeland, Familias 3 – extensions and new functionality, Forensic Sci. Int. Genet. 13 (November) (2014) 121–127, https://doi.org/10.1016/j.fsigen.2014.07.004.
[4] Daniel Kling, Andreas Tillmar, Thore Egeland, Petter Mostad, A general model for likelihood computations of genetic marker data accounting for linkage, linkage disequilibrium, and mutations, Int. J. Leg. Med. 129 (5) (2015) 943–954, https://doi.org/10.1007/S00414-014-1117-7.

2.3. STR mutation rates modeling

To assess the correlation between the sequence of the markers' repetitive motif and the frequency of multistep mutations, allelic transmission data were gathered from published the literature. For this, data from simple structured Y-STRs were used, as markers from the MSY are ideal for the study and modeling of STR mutation rates since the identification of the paternal and filial alleles is unequivocal.

This work resulted in a submission to Scientific Reports in June 2022 that is at the time of writing on the second round of revisions. It is presented in section 2.3.1. of this thesis.

2.3.1. The sequence of the repetitive motif influences the frequency of multistep mutations in Short Tandem Repeats

In revision process for Scientific Reports. For supplementary material please check Appendix F.

Sofia Antão-Sousa[1,2,3,4,*], Nádia Pinto[1,2,5], Pablo Rende[1,3], António Amorim[1,2,3], Leonor Gusmão[4]

[1] Institute for Research and Innovation in Health Sciences (i3S), University of Porto, Porto Portugal.
[2] Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal.
[3] Department of Biology, Faculty of Sciences of University of Porto (FCUP), Porto, Portugal.
[4] DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Rio de Janeiro, Brazil.
[5] Center of Mathematics of University of Porto (CMUP), Porto, Portugal.
* Email address: antaosousasofia@gmail.com

Abstract

Microsatellites, or Short Tandem Repeats (STRs), are subject to frequent length mutations that involve the loss or gain of an integer number of repeats. This work aimed to investigate the correlation between STRs' specific repetitive motif composition and mutational dynamics, specifically the occurrence of single- or multistep mutations. Allelic transmission data, comprising 321,244 allele transfers and 1,297 mutations, were gathered for 37 Y-chromosomal STRs with simple structure. Six structure groups were established: ATT, CTT, TCTA/GATA, GAAA/CTTT, CTTTT, and AGAGAT, according to the repetitive motif present in the DNA leading strand of the markers. Results show that the occurrence of multistep mutations varies significantly among groups of markers defined by the repetitive motif. The group of markers with the highest frequency of multistep mutations was the one with repetitive motif CTTTT (25% of the detected mutations) and the lowest frequency corresponding to the group with repetitive motifs TCTA/GATA (0.93%). Statistically significant differences (α=0.05) were found between groups with repetitive motifs with different lengths, as is the case of TCTA/GATA and ATT ($p = 0.0168$), CTT ($p < 0.0001$), and CTTTT ($p < 0.0001$), as well as between GAAA/CTTT and CTTTT ($p = 0.0102$). The same occurred between the two tetrameric

groups GAAA/CTTT and TCTA/GATA ($p < 0.0001$) – the first showing 5.7 times more multistep mutations than the second. When considering the number of repeats of the mutated paternal alleles, statistically significant differences were found for alleles with 10 or 12 repeats, between GATA and ATT structure groups. These results, which demonstrate the heterogeneity of mutational dynamics across repeat motifs, have implications in the fields of population genetics, epidemiology, or phylogeography, and whenever STR mutation models are used in evolutionary studies in general.

Keywords: microsatellite, STR, length mutations, single-step mutations, multistep mutations, stepwise mutation model, evolution.

Introduction

Microsatellites, or short tandem repeats (STRs), consist of tandemly arrayed 1-6 base pairs (bp) motifs. These are among the most useful and commonly employed genetic markers in population, forensic, or conservation genetics (Ellegren, 2004), due to their variability and ubiquity. Their instability has relevant medical implications, being linked to cancer (Panzer *et al.*, 1995) and to many other diseases. Namely, there are over 40 neurological, neurodegenerative, and neuromuscular disorders determined by repeat expansions of STRs at coding and non-coding regions (Pearson *et al.*, 2005).

STRs undergo rapid length changes due to the insertion or deletion of one or multiple repeat units (Ellegren, 2004; Pearson *et al.*, 2005). The primary mutational mechanism thought to lead to changes in STR length is polymerase template slippage during DNA replication (Schlötterer & Tautz, 1992; Strand *et al.*, 1993). A distinct pathway is associated with unequal crossing over, which may happen due to strand mispairing during recombination (Eckert & Hile, 2009).

The stepwise mutation model (SMM) was introduced by Ohta and Kimura (Ohta & Kimura, 1973) and Wehrhahn (Wehrhahn, 1975), suggesting mutational dynamics of STRs where parental alleles gain or lose a single repeat when transmitted to the offspring. The possibility of multistep changes was also considered, although at a much lower rate. Indeed, some works showed that the proportion of multistep mutations represents 1% of the detected mutations for tri- and tetranucleotide STRs, increasing this figure to 30% for dinucleotide STRs (Gymrek *et al.*, 2017; J. X. Sun *et al.*, 2012). The SMM has been used to model STR mutation and evolution and has been applied in diverse areas such as population genetics (H. Xu & Fu, 2004), epidemiology (Aandahl

et al., 2012), or phylogeography (Barthe *et al.*, 2012). The traditional approach for quantifying kinship likelihood ratios relies on establishing a value corresponding to the decreased probability for each additional repeat difference between parental and filial alleles. This so-called "mutation range" parameter is considered in diverse software (*Forensic Mathematics*, n.d.; Kling *et al.*, 2014). Despite the lack of statistical support, 0.1 is sometimes suggested as an overall value for the mutation range, meaning that a two-step mutation is 10 times rarer than a single-step one, and a three-step mutation is 10 times rarer than a two-step one, and so on.

To investigate the impact of the composition of the STR's repetitive motif in the mutational dynamics, we have compiled data available for STRs located in the non-recombining region of the Y chromosome (Y-STRs). This region of the Y chromosome possesses no homologous in the X chromosome and, as such, they do not undergo recombination during meiosis. Hence, in markers with simple sequence, any change detected between father and son must be due to a mutation event. It is also noteworthy that the data obtained for this study were generated through genotyping platforms that do not discriminate variation in sequence (automated fragment size determination after capillary electrophoresis), but just differences in alleles' length.

Indeed, the Y chromosome is an invaluable tool for the study of germinal mutations and their biological mechanisms since it is exclusively transmitted through the paternal lineage in a haploid fashion. The NRY accommodates many STRs. When typing platforms discriminate solely the length of the allele, the Y chromosome, due to its specific mode of transmission, is the only component of the nuclear genome that allows the exact knowledge of which parental allele resulted in which filial one, allowing the unambiguous identification of any length mutation (Pinto *et al.*, 2014b).

In both autosomal and heterosomal modes of transmission, when no Mendelian incompatibilities are detected in parent(s)-child duos or trios, it is assumed that no mutation occurred. This unavoidably leads to an underestimation of the mutation rates, since 'hidden' or 'covert' mutations may be present (Brenner, 2004; Chakraborty *et al.*, 1996; Vicard & Dawid, 2004).

A most parsimonious approach is used when classifying the mutation as either single- or multistep, i.e., the mutation that requires the minimum number of steps to conciliate the observations with Mendelian transmission is assumed. This leads to an overestimation of the single-step mutation rates and a corresponding underestimation of those involving multiple steps. It is however noteworthy that this is more severe for autosomal than for X-chromosomal markers, since in father-daughter and mother-son

transmissions the parental and filial alleles, respectively, are known (Antão-Sousa, Conde-Sousa, *et al.*, 2019).

Here we intend to contribute to the improvement of the estimates and the mutational model design, by correlating the Y chromosome-specific STRs (Y-STRs) repetitive motif sequence, rather than just its length, with the mutational dynamics.

We have found that the frequency of multistep mutations varies widely across repeat motif compositions and length, reaching differences by a factor of nearly an order of magnitude. The implications of these findings in the fields of population genetics, epidemiology, or phylogeography, and in general evolutionary studies where STR mutation models are discussed.

Material and Methods

Data from 44 published reports (Adnan *et al.*, 2018; Ambrosio *et al.*, 2020; Antão-Sousa *et al.*, 2017; Ay *et al.*, 2019; Ballantyne *et al.*, 2014; Ballard *et al.*, 2005; Berger *et al.*, 2005; Blanchi *et al.*, 1998; Bredemeyer *et al.*, 2021; Budowle *et al.*, 2005; Bugoye *et al.*, 2018; Decker *et al.*, 2008; Domingues *et al.*, 2007; Dupuy *et al.*, 2001, 2004; Fu *et al.*, 2020; Ge *et al.*, 2009; Goedbloed *et al.*, 2009; Gusmão *et al.*, 2005; Hohoff *et al.*, 2007; Kayser *et al.*, 2000; Kim *et al.*, 2009; Kumagai *et al.*, 2007; Kurihara *et al.*, 2004; Laouina *et al.*, 2013; Lee *et al.*, 2007; Lessig & Edelmann, 1998; Lin *et al.*, 2020; Mertoglu *et al.*, 2018; Padilla-Gutiérrez *et al.*, 2008; Petrovic *et al.*, 2019; Pontes *et al.*, 2007; Robino *et al.*, 2015; Sánchez-Diz *et al.*, 2008; Shi *et al.*, 2007; Soares-Vieira *et al.*, 2008; Tsai *et al.*, 2002; Turrina *et al.*, 2006; Vieira-Silva *et al.*, 2009; Wang *et al.*, 2016; Wu *et al.*, 2018; Yang *et al.*, 2018; Yuan *et al.*, 2019; Zhang *et al.*, 2019) were gathered, comprising a total of 2,320 observed mutations in 451,003 allele transfers between father and son pairs, regarding 64 Y-STRs (see **Table S1 and S2**). These data were obtained in genotyping platforms (automated fragment size determination after capillary electrophoresis) that do not discriminate variation in sequence, but just size differences. As previously referred, a change between a pair of paternal and filial Y-STR alleles implies that a mutation occurred, but the inverse is only true if the analyzed markers have simple structure. Also, only using STRs with simple structure is possible to determine the number of repeats involved in the allelic transmission. Thus, after compiling data from all studies including father-child duos, DYS389II, DYS390, DYS435, DYS447, DYS520, DYS547, DXY156 were excluded from the analyses because they harbor complex structures. Markers containing several loci (multi-copy), such as DYS385a/b, DYS459a/b, DYS464a/b/c/d, DYS526a/b, DYS527a/b, DYF387S1, DYF399S1,

DYF404S1, DYF403S1a/b, were also not considered since they do not allow the unambiguous assignment of mutation to each locus. Structure groups with fewer than 10 reported mutations were also removed from the analyses: DYS531 and DYS587, due to a lack of statistical power. Finally, DYS622, DYS630 and DYS640 were not considered since no sequence information was found.

A final subset of 37 Y-STRs, 321,244 allele transfers and 1,297 mutations, was then considered for further analyses (see **Table S1**).

STRs were grouped according to the sequence and length of the repetitive motif present in the leading strand (retrieved from GRCh38.p14 (*GRCh38.P14 - Hg38 - Genome - Assembly - NCBI*, n.d.)), resulting in 8 groups, as shown in **Table 1**.

In forensic genetics, STRs nomenclature recommendations state that, although most times it is possible to define different repetitive motifs within a 5' to 3' strand, the repeat sequence motif must be defined so that the first 5'-nucleotides that can represent a repetitive motif are used (Bar *et al.*, 1997). However, when a mutation occurs, it is impossible to discern if the length change resulted in the addition or deletion of the designated repetitive motif or any other. For example, if the repetitive motif of an STR is defined as TCTA, when a length mutation occurs, that repetitive motif might have been the one involved in the mutation, but so could the motifs CTAT, TATC, and ATCT (see **Table 1** for the group information). It is impossible to discern which motif was involved in the mutation through capillary electrophoresis or sequencing. As such, in this work, STRs were grouped according to their structure and not their official nomenclature.

Table 1: Grouping of the STRs analyzed according to the repetitive motif present in the leading strand.

Group	Structure groups	Markers
ATT	**AAT/ATA/TAA**	DYS388 DYS392
CTT	**CTT/TCT/TTC**	DYS481 DYS612
GATA	**GATA/AGAT/TAGA/ATAG**	DYS393 DYS456 DYS522 DYS635 DYS439 DYS389I DYS549

Group	Structure groups	Markers
		DYS444
		DYS510
TCTA	**TCTA/CTAT/TATC/ATCT**	DYS19
		DYS391
		DYS437
		DYS533
		GATA H4
		DYS460
		DYS461
		GATA A10
		DYS434
		DYS552
GAAA	**GAAA/AGAA/AAGA/AAAG**	DYS576
		DYS627
		DYS518
		DYS458
		DYS626
		DYS722
CTTT	**CTTT/TCTT/TTCT/TTTC**	DYS570
		DYS449
		DYS557
		DYS709
CTTTT	**CTTTT/TCTTT/TTCTT/TTTCT/TTTTC**	DYS643
		DYS446
		DYS438
AGAGAT	**AGAGAT**	DYS448

As TCTA and GATA, and GAAA and CTTT are complementary sequences, to determine if they could be grouped, Fisher exact tests were performed to ascertain the statistical significance of the differences in the number of single- and multistep mutations between the two pairs (α=0.05). No significant differences were detected in the comparison of GAAA with CTTT markers ($p = 0.8415$) nor in the comparison of TCTA with GATA markers ($p = 0.0846$). Hence, GAAA were grouped with CTTT markers, and GATA were grouped with TCTA markers.

The ratio between single- and multistep mutations was calculated for each of the above-defined groups of markers. Fisher's exact tests were also used to measure the significance (α=0.05) of the single/multistep proportions between groups of markers.

The number of repeats involved in allele transitions where mutations were observed was also analyzed for the complete set of 37 single-copy Y-STRs with simple structure.

In markers DYS19, DYS389I, and DYS635, allele calling includes the total number of repeats in polymorphic and contiguous non-polymorphic tracts. Proper adjustments were made for these markers to obtain the number of repeats of the polymorphic tract.

Some of the published reports [[51], [56], [58], [61], [63]] do not indicate the alleles observed in the mutation, providing only information on the type of mutation observed (single- or multistep, gain or loss of repeats). These works were thus not included in the analyses involving the number of repeats.

Results and discussion

Although many studies report single-step mutations as much more frequent than multistep mutations, these results are usually presented as an overall value, and not analyzed per marker - see for example (Decker *et al.*, 2008; Dupuy *et al.*, 2004; Kayser *et al.*, 2000). Our results regarding markers with simple structure show that, indeed, single-step mutations are more frequent than multistep ones (except for marker DYS438, see **Table S1**). However, the ratio between single- and multistep mutations varies widely between markers and groups of markers defined by their repetitive motif structure (see **Table 2**).

The CTTTT group showed the highest frequency of multistep mutations (25% of the mutations observed), more than twice the corresponding frequency of the ATT and CTT groups, with the second-highest frequency (~12%). The lowest frequency of multistep mutations was observed for the group TCTA/GATA (~0.93%).

Comparing the two tetrameric groups, the GAAA/CTTT group showed 5.7 times higher multistep mutation frequency than the group GATA/TCTA, the corresponding confidence intervals not intersecting each other.

Table 2: Number of multistep (a) and total (b) mutations observed, multistep mutation frequency, and corresponding 95% confidence intervals, per group of markers.

Group	Multistep mutations (a)	Mutations (b)	Multistep mutation frequency*	IC 95% lower bound	IC 95% upper bound
ATT	2	17	0.1176	0.0146	0.3644
CTT	10	86	0.1162	0.0572	0.2035
TCTA/ GATA	5	540	0.0093	0.0030	0.0215
GAAA/ CTTT	33	629	0.0524	0.0364	0.0729
CTTTT	4	16	0.2500	0.0727	0.5238
AGAGAT	1	19	0.0526	0.0013	0.2603

* Calculated as: $\frac{a}{b}$. Values rounded to 4 dp.

Ballantyne *et al.* (Ballantyne *et al.*, 2010) concluded that motifs with strong purine:pyrimidine asymmetries have the highest diversity and variance. Our results indicate that this could also be a factor affecting the type of mutation, with a consequent impact on the variance in the number of repeats. For STRs with tetrameric motifs, the GAAA group, with a 4:0 ratio of purine:pyrimidine, presents a greater frequency of multistep mutations than the GATA group, with a 3:1 ratio ($p < 0.0001$, see **Table 3**). The same trend is observed for trimeric repeats, with the CTT having a higher ratio of multistep mutations than the ATT motifs. The frequency of multistep mutations is even higher regarding the pentameric motif CTTTT, with 0:5 ratio of purine:pyrimidine. However, in this case, we cannot discern if this difference is influenced by the higher asymmetry or the larger number of nucleotides in the motif. Significant differences were also found between both ATT and CTT groups and the TCTA/GATA ($p = 0.0168$ and $p < 0.0001$, respectively), and between both TCTA/GATA and GAAA/CTTT groups and the CTTTT ($p < 0.0001$, and $p = 0.0102$, respectively) – see **Table 3**.

Table 3: P-values resulting from a pairwise Fisher test of the number of single- and multistep mutations between the STR groups defined by the repetitive motif (α = 0.05).

	ATT	CTT	TCTA/GATA	GAAA/CTTT	CTTTT	AGAGAT
ATT	-	1.0000	0.0168	0.2336	0.3983	0.5929
CTT	-	-	< 0.0001	0.0281	0.2268	0.6844
TCTA/GATA	-	-	-	< 0.0001	< 0.0001	0.1881
GAAA/CTTT	-	-	-	-	0.0102	1.0000
CTTTT	-	-	-	-	-	0.1558
AGAGAT	-	-	-	-	-	-

Significant p-values underlined, values presented with 4 dp for non-null approximate values, in which case a maximum value with one significant digit is shown.

The correlation between the length of repetitive motif and the mutation rate have been shown in different studies (e.g., (Ballantyne *et al.*, 2010; Chakraborty *et al.*, 1997; Pemberton *et al.*, 2009; Webster *et al.*, 2002)). Most of these studies also acknowledge the presence of mutations that escape SMM, however, without relating their frequency with the repetitive structure of the locus. Beyond these analyses, our work shows how frequently some STRs can escape the SMM. Most mutations obey the SMM, but some escape this model, for some markers and/or groups of markers more than for others. So, despite being the most used model, and suitable for most STRs, the SMM should be used with caution for others.

Martins *et al.* (Martins *et al.*, 2006) found that wild-type Machado-Joseph Disease alleles do not follow the single-step mutation model. Their results show that the frequency distribution of CAG alleles has been shaped by a multistep mutation mechanism. Indeed, this seems to be the case for some of the groups in this work, that show multistep mutation proportions up to 25%.

Most works show a considerable disproportion between single- and multistep mutations, which might be due to the high number of GATA markers analyzed in the most used multiplexes. In the last years, more GAAA markers have been added to the commercially available typing kits and the ratio between single- and multistep mutations will likely tend to be less disproportionate. Penta and hexameric motifs are much less represented in the generally used commercial kits and so their effect on these overall rates has little impact.

The number of single- and multistep mutations considering the number of repeats involved in the allele transmissions were analyzed for the complete set of markers and structure groups – see Table 4.

Table 4: Number of single- and multistep mutations considering the number of repeats of the parental alleles for all the structure groups.

Number of repeats	Structure Group											
	CTAT/ GATA		GAAA/ CTTT		ATT		CTT		GAAAA		AGAGAT	
	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi
6	1	0										
7	1	0										
8	3	0										
9	17	0										
10*	43	0			0	1			4	1		
11	112	3			2	0						
12**	86	0			1	0			1	3		
13	85	0			6	0						
14	49	0	1	0	2	1						
15	41	0	10	0								
16	24	0	20	1								
17	15	1	62	5								
18	3	0	69	2							2	0
19§	0	1	57	4							7	0
19.2			1	0								
20			26	5			1	0			3	0
21			36	1			3	0			1	1
22			23	1			2	1			1	0
23			15	0			4	1				
24			7	0			16	1				
25			3	0			7	2				
26			2	0			15	0				
27			3	1			10	3				
28			2	0			5	0				
29			10	0			0	0				
30			11	0			0	0				
31§§			15	0			0	1				
32			21	0								
33			20	1								
34			11	0								
35			10	1								
36			3	0								
37			8	1								

Number of repeats	Structure Group											
	CTAT/ GATA		GAAA/ CTTT		ATT		CTT		GAAAA		AGAGAT	
	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi
38			12	2								
39			15	0								
40			11	2								
41			10	1								
42			9	0								
43			5	0								
44			1	1								
45			0	1								

* to ** significant p-values found (α=0.05), $^{*}p = 0.0227$ for CTAT/GATA and ATT, $^{**}p < 0.0001$ for CTAT/GATA and GAAAA

§ to §§ nearly significant p-values found (α=0.05), $^{\S}p = 0.0806$ for CTAT/GATA and GAAA/CTTT, $^{\S\S}p = 0.0625$ for GAAA/CTTT and CTT

The high number of categories considered through this approach implies a low number of observations in each of them. This implies that differences may not be detected even if they exist. Nevertheless, for a set of 22 numbers of repeats existing in at least 2 structure groups, 2 showed statistically significant differences (and 2 nearly significant). This supports that, at least in some cases, the structure of the repeat motif does influence the proportion of single- and multistep mutation, beyond the length of the polymorphic tract.

Conclusions

So far, diverse studies have shown the influence of several factors on STRs mutation rates, such as the allele length, repeat motif size and sequence, parental sex, and age. Others have studied the correlation between the mutation rate and the nucleotide composition of the repetitive motif with the same number of base pairs (see, for example, (Ballantyne *et al.*, 2010)). However, the influence of nucleotide composition of the repetitive motif on the type of mutation (single- or multistep), was not systematically investigated. In this study, we took advantage of the mode of transmission of the non-recombining region of the Y chromosome, which enables the direct analysis of length mutations in markers with simple structure.

Despite the inescapable problem regarding the low number of observations when modeling rare events, this work supports that, just like mutation rates, the type of mutation (single- or multistep) is heterogeneous across STRs. This includes markers with the same length of the repetitive motif, as well as alleles with the same number of repeats, although from different markers. Comparing repetitive motifs with different sizes prevent us to discern the reason leading to the observed unbalance between single- and multistep mutations. In any case, our work supports that the best fitting mutation model varies between markers.

The monomeric tract in motifs ATT, CTT, GAAA and CTTTT might be influencing slippage, or another mutation model might be operating since in these motifs the multistep mutation frequency is higher.

Most noteworthy, is the case of one of the pentameric markers analyzed, DYS438, which does not fit the single-step mutation model, as half of the observed mutations involved several steps.

It is clear that, at least for some STR motif structures, the single stepwise mutation model represents, at best, a crude and biased oversimplification. The implications are manifold and affect many areas of study, such as human population and evolutionary history, genealogical studies, or forensics. Concerning forensic applications, the "mutation range" parameter of 0.1 frequently used in kinship computations seems to be too high for all tetrameric STRs analyzed and too low for pentameric ones. Based on the available data, the mutation range parameter estimates are 0.1333 for ATT markers, 0.1316 for CTT, 0.0554 for GAAA/CTTT, 0.0093 for motif TCTA/GATA, 0.3333 for CTTTT, and 0.0556 for AGAGAT (although in these two last cases more data are needed for a sound estimate).

The development of new models of STR evolution including all major factors known to influence mutation is challenging, but their development is crucial. Large datasets are needed to test mutation models and to estimate rates more accurately. One major setback is that some markers have extremely low mutation rates and gathering enough data is challenging, in such cases targeted analyses are needed. Moreover, guidelines concerning mutation reporting should be established, a need particularly felt when dealing with STRs outside NRY, as previously mentioned in (Amorim & Pinto, 2020). These data should include parental age, and genotypic information, as the absolute frequencies of the observed alleles in one-generation profiles (separately for duos and trios in the case of either autosomal or X chromosomal markers, comprising all the cases, with or without mutation, and for the full set of analyzed markers). Such enriched and organized datasets would improve mutation modeling, enabling allele-specific mutation

rates estimates, and allowing the discernment and quantification of the effects of the various factors influencing the fidelity of the genetic transmission.

Acknowledgments

This work was partially financed by FEDER- Fundo Europeu de Desenvolvimento Regional funds through the COMPETE 2020 - Operational Program for Competitiveness and Internationalization (POCI), Portugal 2020, and by Portuguese funds through FCT-Fundação para a Ciência e a Tecnologia/Ministério da Ciência, Tecnologia e Inovação in the framework of the projects "Institute for Research and Innovation in Health Sciences" (POCI-01-0145-FEDER-007274). NP is supported by FCT, under the program contract provided in Decree-Law no.57/2016 of August 29. SAS is funded by a FCT doctoral grant SFRH/BD/136284/2018. LG is supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (ref. 306342/2019-7), and Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro - FAPERJ (CNE-2022).

References

[1] H. Ellegren, "Microsatellites: simple sequences with complex evolution," *Nature Reviews Genetics 2004 5:6*, vol. 5, no. 6, pp. 435–445, Jun. 2004, doi: 10.1038/nrg1348.

[2] S. Panzer, D. P. A. Kuhl, and C. T. Caskey, "Unstable triplet repeat sequences: A source of cancer mutations?," *Stem Cells*, vol. 13, no. 2, pp. 146–157, Jan. 1995, doi: 10.1002/STEM.5530130206.

[3] C. E. Pearson, K. N. Edamura, and J. D. Cleary, "Repeat instability: mechanisms of dynamic mutations," *Nature Reviews Genetics 2005 6:10*, vol. 6, no. 10, pp. 729–742, Oct. 2005, doi: 10.1038/nrg1689.

[4] C. Schlötterer and D. Tautz, "Slippage synthesis of simple sequence DNA.," *Nucleic Acids Res*, vol. 20, no. 2, p. 211, Jan. 1992, doi: 10.1093/NAR/20.2.211.

[5] M. Strand, T. A. Prolla, R. M. Liskay, and T. D. Petes, "Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair," *Nature 1993 365:6443*, vol. 365, no. 6443, pp. 274–276, 1993, doi: 10.1038/365274a0.

[6] K. A. Eckert and S. E. Hile, "Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome," *Mol Carcinog*, vol. 48, no. 4, pp. 379–388, Apr. 2009, doi: 10.1002/MC.20499.

[7] T. Ohta and M. Kimura, "A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population," *Genet Res*, vol. 22, no. 2, pp. 201–204, 1973, doi: 10.1017/S0016672300012994.

[8] C. F. Wehrhahn, "The evolution of selectively similar electrophoretically detectable alleles in finite natural populations," *Genetics*, vol. 80, no. 2, pp. 375–394, 1975, doi: 10.1093/GENETICS/80.2.375.

[9] M. Gymrek, T. Willems, D. Reich, and Y. Erlich, "Interpreting short tandem repeat variations in humans using mutational constraint," *Nature Genetics 2017 49:10*, vol. 49, no. 10, pp. 1495–1501, Sep. 2017, doi: 10.1038/ng.3952.

[10] J. X. Sun *et al.*, "A direct characterization of human mutation based on microsatellites," *Nat Genet*, vol. 44, no. 10, pp. 1161–1165, Oct. 2012, doi: 10.1038/NG.2398.

[11] H. Xu and Y. X. Fu, "Estimating Effective Population Size or Mutation Rate With Microsatellites," *Genetics*, vol. 166, no. 1, pp. 555–563, Jan. 2004, doi: 10.1534/GENETICS.166.1.555.

[12] R. Z. Aandahl, J. F. Reyes, S. A. Sisson, and M. M. Tanaka, "A Model-Based Bayesian Estimation of the Rate of Evolution of VNTR Loci in Mycobacterium tuberculosis," *PLoS Comput Biol*, vol. 8, no. 6, p. e1002573, Jun. 2012, doi: 10.1371/JOURNAL.PCBI.1002573.

[13] S. Barthe, F. Gugerli, N. A. Barkley, L. Maggia, C. Cardi, and I. Scotti, "Always Look on Both Sides: Phylogenetic Information Conveyed by Simple Sequence Repeat Allele Sequences," *PLoS One*, vol. 7, no. 7, p. e40699, Jul. 2012, doi: 10.1371/JOURNAL.PONE.0040699.

[14] "Forensic Mathematics." https://dna-view.com/#Software (accessed Apr. 28, 2022).

[15] D. Kling, A. O. Tillmar, and T. Egeland, "Familias 3 – Extensions and new functionality," *Forensic Sci Int Genet*, vol. 13, pp. 121–127, Nov. 2014, doi: 10.1016/j.fsigen.2014.07.004.

[16] N. Pinto, L. Gusmão, and A. Amorim, "Mutation and mutation rates at y chromosome specific Short Tandem Repeat Polymorphisms (STRs): A reappraisal," *Forensic Sci Int Genet*, vol. 9, no. 1, pp. 20–24, Mar. 2014, doi: 10.1016/j.fsigen.2013.10.008.

[17] R. Chakraborty, D. N. Stivers, and Y. Zhong, "Estimation of mutation rates from parentage exclusion data: applications to STR and VNTR loci," *Mutat Res*, vol. 354, no. 1, pp. 41–48, Jul. 1996, doi: 10.1016/0027-5107(96)00014-0.

[18] P. Vicard and A. P. Dawid, "A statistical treatment of biases affecting the estimation of mutation rates," *Mutat Res*, vol. 547, no. 1–2, pp. 19–33, Mar. 2004, doi: 10.1016/J.MRFMMM.2003.11.005.

[19] C. H. Brenner, "Multiple mutations, covert mutations and false exclusions in paternity casework," *Int Congr Ser*, vol. 1261, no. C, pp. 112–114, Apr. 2004, doi: 10.1016/S0531-5131(03)01843-0.

[20] S. Antão-Sousa, E. Conde-Sousa, L. Gusmão, A. Amorim, and N. Pinto, "Underestimation and misclassification of mutations at X chromosome STRs depend on population's allelic profile," *Forensic Sci Int Genet Suppl Ser*, vol. 7, no. 1, pp. 718–720, Dec. 2019, doi: 10.1016/J.FSIGSS.2019.10.150.

[21] R. Lessig and J. Edelmann, "Y chromosome polymorphisms and haplotypes in west Saxony (Germany)," *Int J Legal Med*, vol. 111, no. 4, pp. 215–218, Jun. 1998, doi: 10.1007/S004140050155.

[22] N. O. Blanchi *et al.*, "Characterization of Ancestral and Derived Y-Chromosome Haplotypes of New World Native Populations," *The American Journal of Human Genetics*, vol. 63, no. 6, pp. 1862–1871, Dec. 1998, doi: 10.1086/302141.

[23] M. Kayser *et al.*, "Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs.," *Am J Hum Genet*, vol. 66, no. 5, p. 1580, 2000, doi: 10.1086/302905.

[24] B. M. Dupuy, M. Stenersen, T. Egeland, and B. Olaisen, "Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci," *Hum Mutat*, vol. 23, no. 2, pp. 117–124, 2004, doi: 10.1002/HUMU.10294.

[25] L. C. Tsai *et al.*, "Haplotype frequencies of nine Y-chromosome STR loci in the Taiwanese Han population," *Int J Legal Med*, vol. 116, no. 3, pp. 179–183, 2002, doi: 10.1007/S004140100236.

[26] R. Kurihara *et al.*, "Mutations in 14 Y-STR loci among Japanese father-son haplotypes," *Int J Legal Med*, vol. 118, no. 3, pp. 125–131, Jun. 2004, doi: 10.1007/S00414-003-0422-3.

[27] B. Budowle *et al.*, "Twelve short tandem repeat loci Y chromosome haplotypes: genetic analysis on populations residing in North America," *Forensic Sci Int*, vol. 150, no. 1, pp. 1–15, May 2005, doi: 10.1016/J.FORSCIINT.2005.01.010.

[28] B. Berger, A. Lindinger, H. Niederstätter, P. Grubwieser, and W. Parson, "Y-STR typing of an Austrian population sample using a 17-loci multiplex PCR assay," *Int J Legal Med*, vol. 119, no. 4, pp. 241–246, 2005, doi: 10.1007/S00414-005-0546-8.

[29] L. Gusmão *et al.*, "Mutation rates at Y chromosome specific microsatellites," *Hum Mutat*, vol. 26, no. 6, pp. 520–528, Dec. 2005, doi: 10.1002/HUMU.20254.

[30] D. J. Ballard *et al.*, "A study of mutation rates and the characterisation of intermediate, null and duplicated alleles for 13 Y chromosome STRs," *Forensic Sci Int*, vol. 155, no. 1, pp. 65–70, Dec. 2005, doi: 10.1016/J.FORSCIINT.2004.12.012.

[31] S. Turrina, R. Atzei, and D. de Leo, "Y-chromosomal STR haplotypes in a Northeast Italian population sample using 17plex loci PCR assay," *Int J Legal Med*, vol. 120, no. 1, pp. 56–59, Jan. 2006, doi: 10.1007/S00414-005-0054-X.

[32] H. Y. Lee *et al.*, "Haplotypes and mutation analysis of 22 Y-chromosomal STRs in Korean father-son pairs," *Int J Legal Med*, vol. 121, no. 2, pp. 128–135, Mar. 2007, doi: 10.1007/S00414-006-0130-X.

[33] C. Hohoff, K. Dewa, U. Sibbing, K. Hoppe, P. Forster, and B. Brinkmann, "Y-chromosomal microsatellite mutation rates in a population sample from northwestern Germany," *Int J Legal Med*, vol. 121, no. 5, pp. 359–363, Sep. 2007, doi: 10.1007/S00414-006-0123-9.

[34] R. Kumagai, A. Kumagai, K. Saigusa, and Y. Aoki, "Haplotype analysis of 17 Y-STR loci in a Japanese population," *Forensic Sci Int*, vol. 172, no. 1, pp. 72–78, Oct. 2007, doi: 10.1016/J.FORSCIINT.2006.07.022.

[35] M. L. Pontes, L. Cainé, D. Abrantes, G. Lima, and M. F. Pinheiro, "Allele frequencies and population data for 17 Y-STR loci (AmpFlSTR Y-filer) in a Northern Portuguese population sample," *Forensic Sci Int*, vol. 170, no. 1, pp. 62–67, Jul. 2007, doi: 10.1016/J.FORSCIINT.2006.04.008.

[36] M. sen Shi, J. P. Tang, R. F. Bai, X. J. Yu, J. Y. Lv, and B. Hu, "Haplotypes of 20 Y-chromosomal STRs in a population sample from southeast China (Chaoshan area)," *Int J Legal Med*, vol. 121, no. 6, pp. 455–462, Nov. 2007, doi: 10.1007/S00414-007-0186-2.

[37] P. M. Domingues, L. Gusmão, D. A. da Silva, A. Amorim, R. W. Pereira, and E. F. de Carvalho, "Sub-Saharan Africa descendents in Rio de Janeiro (Brazil): population and mutational data for 12 Y-STR loci," *Int J Legal Med*, vol. 121, no. 3, pp. 238–241, May 2007, doi: 10.1007/S00414-007-0154-X.

[38] A. E. Decker, M. C. Kline, J. W. Redman, T. M. Reid, and J. M. Butler, "Analysis of mutations in father-son pairs with 17 Y-STR loci," *Forensic Sci Int Genet*, vol. 2, no. 3, Jun. 2008, doi: 10.1016/J.FSIGEN.2007.08.016.

[39] P. Sánchez-Diz *et al.*, "Population and segregation data on 17 Y-STRs: results of a GEP-ISFG collaborative study," *Int J Legal Med*, vol. 122, no. 6, pp. 529–533, Nov. 2008, doi: 10.1007/S00414-008-0265-Z.

[40] J. R. Padilla-Gutiérrez *et al.*, "Population data and mutation rate of nine Y-STRs in a mestizo Mexican population from Guadalajara, Jalisco, México," *Leg Med (Tokyo)*, vol. 10, no. 6, pp. 319–320, Nov. 2008, doi: 10.1016/J.LEGALMED.2008.03.004.

[41] M. Goedbloed *et al.*, "Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFlSTR Yfiler PCR amplification kit," *Int J Legal Med*, vol. 123, no. 6, pp. 471–482, Nov. 2009, doi: 10.1007/S00414-009-0342-Y.

[42] J. Ge, B. Budowle, X. G. Aranda, J. v. Planz, A. J. Eisenberg, and R. Chakraborty, "Mutation rates at Y chromosome short tandem repeats in Texas populations," *Forensic Sci Int Genet*, vol. 3, no. 3, pp. 179–184, Jun. 2009, doi: 10.1016/J.FSIGEN.2009.01.007.

[43] S. H. Kim *et al.*, "Population genetics and mutational events at 6 Y-STRs in Korean population," *Forensic Sci Int Genet*, vol. 3, no. 2, Mar. 2009, doi: 10.1016/J.FSIGEN.2008.06.004.

[44] K. N. Ballantyne *et al.*, "Toward male individualization with rapidly mutating y-chromosomal short tandem repeats," *Hum Mutat*, vol. 35, no. 8, pp. 1021–1032, 2014, doi: 10.1002/HUMU.22599.

[45] C. Robino *et al.*, "Development of an Italian RM Y-STR haplotype database: Results of the 2013 GEFI collaborative exercise," *Forensic Sci Int Genet*, vol. 15, pp. 56–63, 2015, doi: 10.1016/J.FSIGEN.2014.10.008.

[46] Y. Wang *et al.*, "Genetic polymorphisms and mutation rates of 27 Y-chromosomal STRs in a Han population from Guangdong Province, Southern China," *Forensic Sci Int Genet*, vol. 21, pp. 5–9, Mar. 2016, doi: 10.1016/J.FSIGEN.2015.09.013.

[47] S. Antão-Sousa *et al.*, "Mutation rates and segregation data on 16 Y-STRs: An update to previous GHEP-ISFG studies," *Forensic Sci Int Genet Suppl Ser*, vol. 6, pp. e601–e602, Dec. 2017, doi: 10.1016/J.FSIGSS.2017.10.008.

[48] A. Adnan, A. Rakha, O. Lao, and M. Kayser, "Mutation analysis at 17 Y-STR loci (Yfiler) in father-son pairs of male pedigrees from Pakistan," *Forensic Sci Int Genet*, vol. 36, pp. e17–e18, Sep. 2018, doi: 10.1016/J.FSIGEN.2018.07.001.

[49] F. C. Bugoye, E. Mulima, and G. Misinzo, "Analysis of Mutation Rate of 17 Y-Chromosome Short Tandem Repeats Loci Using Tanzanian Father-Son Paired Samples," *Genet Res Int*, vol. 2018, 2018, doi: 10.1155/2018/8090469.

[50] E. Mertoglu, G. Filoglu, T. Zorlu, and O. Bulbul, "Estimation of the Y-chromosomal short tandem repeat (Y-STR) mutation rates in Turkey," *Turkish Journal of Biochemistry*, vol. 43, no. 2, pp. 142–150, Apr. 2018, doi: 10.1515/TJB-2017-0240/PDF.

[51] Y. Yang *et al.*, "Haplotypic polymorphisms and mutation rate estimates of 22 Y-chromosome STRs in the Northern Chinese Han father–son pairs," *Scientific Reports 2018 8:1*, vol. 8, no. 1, pp. 1–6, May 2018, doi: 10.1038/s41598-018-25362-3.

[52] W. Wu *et al.*, "Mutation rates at 42 Y chromosomal short tandem repeats in Chinese Han population in Eastern China," *Int J Legal Med*, vol. 132, no. 5, pp. 1317–1319, Sep. 2018, doi: 10.1007/S00414-018-1784-X.

[53] V. Petrovic, M. Kecmanovic, M. Keckarevic Markovic, and D. Keckarevic, "Assessment of mutation rates for PPY23 Y chromosome STR loci in Serbian father-son pairs," *Forensic Sci Int Genet*, vol. 39, pp. e5–e9, Mar. 2019, doi: 10.1016/J.FSIGEN.2018.11.014.

[54] J. Zhang *et al.*, "Mutation rates in father-son pairs of the 27 Y-STR loci in the Dezhou Han population from Shandong province, eastern China," *J Forensic Leg Med*, vol. 67, pp. 61–63, Oct. 2019, doi: 10.1016/J.JFLM.2019.06.015.

[55] M. Ay, A. Serin, H. Sevay, C. Gurkan, and H. Canan, "Genetic characterisation of 13 rapidly mutating Y-STR loci in 100 father and son pairs from South and East Turkey," *https://doi.org/10.1080/03014460.2018.1559353*, vol. 45, no. 6–8, pp. 506–515, Nov. 2019, doi: 10.1080/03014460.2018.1559353.

[56] L. Yuan *et al.*, "Mutation analysis of 13 RM Y-STR loci in Han population from Beijing of China," *Int J Legal Med*, vol. 133, no. 1, pp. 59–63, Jan. 2019, doi: 10.1007/S00414-018-1949-7.

[57] I. B. Ambrosio *et al.*, "Mutational data and population profiling of 23 Y-STRs in three Brazilian populations," *Forensic Sci Int Genet*, vol. 48, Sep. 2020, doi: 10.1016/J.FSIGEN.2020.102348.

[58] H. Lin, Q. Ye, P. Tang, T. Mo, X. Yu, and J. Tang, "Analyzing genetic polymorphism and mutation of 44 Y-STRs in a Chinese Han population of Southern China," *Leg Med (Tokyo)*, vol. 42, Feb. 2020, doi: 10.1016/J.LEGALMED.2019.101643.

[59] J. Fu, J. Cheng, C. Wei, M. A. Khan, Z. Jin, and J. Fu, "Assessing 23 Y-STR loci mutation rates in Chinese Han father-son pairs from southwestern China," *Mol Biol Rep*, vol. 47, no. 10, pp. 7755–7760, Oct. 2020, doi: 10.1007/S11033-020-05851-3.

[60] S. Bredemeyer, L. Roewer, and S. Willuweit, "Next generation sequencing of Y-STRs in father-son pairs and comparison with traditional capillary electrophoresis," *Forensic Sci Res*, 2021, doi: 10.1080/20961790.2021.1898078/SUPPL_FILE/TFSR_A_1898078_SM7293.XLSX.

[61] C. Vieira-Silva, P. Dario, T. Ribeiro, I. Lucas, H. Geada, and R. Espinheira, "Y-STR mutational rates determination in South Portugal Caucasian population," *Forensic Sci Int*

Genet Suppl Ser, vol. 2, no. 1, pp. 60–61, Dec. 2009, doi: 10.1016/J.FSIGSS.2009.08.159.

[62] J. A. Soares-Vieira, A. E. C. Billerbeck, E. S. M. Iwamura, B. B. Mendonca, L. Gusmão, and P. A. Otto, "Population and mutation analysis of Y-STR loci in a sample from the city of São Paulo (Brazil)," *Genet Mol Biol*, vol. 31, no. 3, pp. 651–656, 2008, doi: 10.1590/S1415-47572008000400009.

[63] A. Laouina *et al.*, "Mutation rate at 17 Y-STR loci in 'Father/Son' pairs from moroccan population," *Leg Med*, vol. 15, no. 5, pp. 269–271, Sep. 2013, doi: 10.1016/J.LEGALMED.2013.03.003.

[64] B. M. Dupuy *et al.*, "Y-chromosome variation in a Norwegian population sample," *Forensic Sci Int*, vol. 117, no. 3, pp. 163–173, Apr. 2001, doi: 10.1016/S0379-0738(00)00397-2.

[65] "GRCh38.p14 - hg38 - Genome - Assembly - NCBI." https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.40 (accessed Apr. 25, 2022).

[66] W. Bar *et al.*, "DNA recommendations. Further report of the DNA Commission of the ISFH regarding the use of short tandem repeat systems. International Society for Forensic Haemogenetics," *Int J Legal Med*, vol. 110, no. 4, pp. 175–176, 1997, doi: 10.1007/S004140050061.

[67] K. N. Ballantyne *et al.*, "Mutability of Y-Chromosomal Microsatellites: Rates, Characteristics, Molecular Bases, and Forensic Implications," *Am J Hum Genet*, vol. 87, no. 3, p. 341, Sep. 2010, doi: 10.1016/J.AJHG.2010.08.006.

[68] R. Chakraborty, M. Kimmel, D. N. Stivers, L. J. Davison, and R. Deka, "Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci," *Proc Natl Acad Sci U S A*, vol. 94, no. 3, pp. 1041–1046, Feb. 1997, doi: 10.1073/PNAS.94.3.1041.

[69] T. J. Pemberton, C. I. Sandefur, M. Jakobsson, and N. A. Rosenberg, "Sequence determinants of human microsatellite variability," *BMC Genomics*, vol. 10, no. 1, pp. 1–19, Dec. 2009, doi: 10.1186/1471-2164-10-612/TABLES/6.

[70] M. T. Webster, N. G. C. Smith, and H. Ellegren, "Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments," *Proc Natl Acad Sci U S A*, vol. 99, no. 13, pp. 8748–8753, Jun. 2002, doi: 10.1073/PNAS.122067599/SUPPL_FILE/0675TABLE6.XLS.

[71] S. Martins, F. Calafell, V. C. Wong, J. Sequeiros, and A. Nio Amorim, "A multistep mutation mechanism drives the evolution of the CAG repeat at MJD/SCA3 locus," *European Journal of Human Genetics*, vol. 14, pp. 932–940, 2006, doi: 10.1038/sj.ejhg.5201643.

[72] A. Amorim and N. Pinto, "Estimates of mutation rates from incompatibilities are misleading - guidelines for publication and retrieval of mutation data urgently needed," 2020, doi: 10.1016/j.fsigss.2019.10.110.

Appendix F

Antão-Sousa, S., Pinto, N., Rende, P., Amorim, A., Gusmão, L.

The sequence of the repetitive motif influences the frequency of multistep mutations in Short Tandem Repeats

Table S3: Number of mutations and allelic transmissions gathered from the 44 published works and the mutational steps for each mutation.

Markers	Number of mutations observed	Number of allelic transmissions	Number of mutational steps						
			1 step	**2 steps**	**3 steps**	**4 steps**	**5 steps**	**6 steps**	**NC***
DYS413 (YCAIII)	1	100	0	1	0	0	0	0	0
YCAII	2	2,296	2	0	0	0	0	0	0
DYS392	14	22,044	13	0	0	1	0	0	0
DYS388	2	3,612	2	0	0	0	0	0	0
DYS481	15	3,390	12	3	0	0	0	0	0
DYS612	66	4,056	60	6	0	0	0	0	0
DYS19	53	24,148	52	0	0	0	0	0	1
DYS389I	62	22,399	62	0	0	0	0	0	0
DYS390	57	22,879	56	1	0	0	0	0	0
DYS391	62	23,567	60	2	0	0	0	0	0
DYS393	29	22,113	29	0	0	0	0	0	0
DYS437	21	18,793	21	0	0	0	0	0	0
DYS439	188	37,774	188	0	0	0	0	0	0
DYS456	66[A]	15,667	64	0	1	0	0	0	0
DYS460	19	4,297	19	0	0	0	0	0	0
DYS461	0	873	0	0	0	0	0	0	0
DYS533	8	4,863	7	1	0	0	0	0	0
DYS549	7	1,811	7	0	0	0	0	0	0
GATA A10	4	1,026	4	0	0	0	0	0	0
GATA H4	40[A]	15,388	38	1	0	0	0	0	0
DYS444	8	2,458	8	0	0	0	0	0	0
DYS522	4	1,954	4	0	0	0	0	0	0
DYS434	0	80	0	0	0	0	0	0	0
DYS510	6	1,160	6	0	0	0	0	0	0
DYS552	9	1,160	9	0	0	0	0	0	0
DYS449	87	7,175	86	0	1	0	0	0	0
DYS458	110	15,695	107	2	1	0	0	0	0
DYS518	97	6,492	87	4	3	2	0	0	1
DYS570	73	7,862	66	6	1	0	0	0	0
DYS576	106	7,763	101	2	2	0	0	0	1
DYS626	41	4,441	39	1	0	0	0	1	0
DYS627	111	6,723	105	5	1	0	0	0	0
DYS385	107	42,217	100	6	1	0	0	0	0

Markers	Number of mutations observed	Number of allelic transmissions	Number of mutational steps						
			1 step	**2 steps**	**3 steps**	**4 steps**	**5 steps**	**6 steps**	**NC***
DYS557	4	1,239	4	0	0	0	0	0	0
DYS709	0	80	0	0	0	0	0	0	0
DYS722	1	640	1	0	0	0	0	0	0
DYF387S1	66	12,590	59	5	1	1	0	0	0
DYF399S1	281	4,053	268	7	3	0	0	0	3
DYF404S1	61[B]	4,052	59	1	0	0	0	0	0
DYS464a/b/c/d	5[B]	473	4	0	0	0	0	0	0
DYS443	2	1,240	2	0	0	0	0	0	0
DYS505	0	64	0	0	0	0	0	0	0
DYS526a	12	3,990	8	3	1	0	0	0	0
DYS435	0	161	0	0	0	0	0	0	0
DYS459a/b	1	2,323	1	0	0	0	0	0	0
DYS531	0	80	0	0	0	0	0	0	0
DYS438	9	18,808	4	3	0	1	0	0	1
DYS643	1	1,172	1	0	0	0	0	0	0
DXY156	0	1,027	0	0	0	0	0	0	0
DYS446	7	1,817	7	0	0	0	0	0	0
DYS587	1	1,160	1	0	0	0	0	0	0
DYS448	19	14,357	18	1	0	0	0	0	0
DYS635 (C4)	50	16,589	50	0	0	0	0	0	0
DYS447	4	2,394	4	0	0	0	0	0	0
DYS389II	83[A]	22,466	80	2	1	0	0	0	0
DYS520	2	1,241	2	0	0	0	0	0	0
DYS526b	47	3,989	45	2	0	0	0	0	0
DYS527	18	3,494	16	2	0	0	0	0	0
DYF403S1a	120	3,889	111	8	1	0	0	0	0
DYF403S1b[C]	36	3,881	30	3	1	1	0	0	0
DYS547	69[B]	4,053	67	0	0	1	0	0	0
DYS622	3	1,240	3	0	0	0	0	0	0
DYS630	10	1,240	10	0	0	0	0	0	0
DYS640	0	640	0	0	0	0	0	0	0

NC* – Non-consensual number of repeats mutation, involving non-integer number of repeats.

[A] – One mutation without allelic information

[B] – Indeterminate/unspecified mutation

[C] – One mutation is a locus deletion in the son

Table S2: Published works where the Y-STRs mutational data were gathered from.

Short Reference	Markers																																	
	DYS19	DYS385	DYS385a	DYS385b	DYS388	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS413 (YCAIII)	DYS434	DYS435	DYS437	DYS438	DYS439	DYS443	DYS444	DYS446	DYS447	DYS448	DYS449	DYS456	DYS458	DYS459a/b	DYS460	DYS461	DYS464a/b/c/d	DYS481	DYS505	DYS510	DYS518
Lessig & Edelmann, 1998 [21]	X	X				X	X	X																										
Bianchi *et al.*, 1998 [22]	X					X	X	X	X	X	X																							
Kayser *et al.*, 2000 [23]	X	X				X	X	X	X	X	X	X																						
Dupuy *et al.*, 2001 [64]	X	X			X	X	X	X	X	X	X																							
Tsai *et al.*, 2002 [25]	X		X	X	X	X	X	X	X	X	X			X																				
Kurihara *et al.*, 2004 [26]	X	X				X	X	X	X	X	X			X	X	X	X										X							
Dupuy *et al.*, 2004 [24]	X	X			X	X	X	X	X	X	X																							
Budowle *et al.*, 2005 [27]	X	X				X	X	X	X	X	X				X	X	X																	
Berger *et al.*, 2005 [9]	X	X				X	X	X	X	X	X				X	X	X					X		X	X									
Gusmão *et al.*, 2005 [29]	X	X				X	X	X	X	X	X				X	X	X										X	X						
Ballard *et al.*, 2005 [30]	X	X				X	X	X	X	X	X				X	X	X										X							
Turrina *et al.*, 2006 [31]	X	X				X	X	X	X	X	X				X	X	X					X		X	X									
Lee *et al.*, 2007 [32]	X	X				X	X	X	X	X	X				X	X	X					X		X	X									
Hohoff *et al.*, 2007 [33]	X	X				X	X	X	X	X	X				X	X	X																	

Short Reference	Markers																																
	DYS19	DYS385	DYS385a	DYS385b	DYS388	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS413 (YCAIII)	DYS434	DYS435	DYS437	DYS438	DYS439	DYS443	DYS444	DYS446	DYS447	DYS448	DYS449	DYS456	DYS458	DYS459a/b	DYS460	DYS461	DYS464a/b/c/d	DYS481	DYS505	DYS510	DYS518
Kumagai *et al.*, 2007 [34]	X	X				X	X	X	X	X	X					X	X							X	X				X				
Pontes *et al.*, 2007 [35]	X		X	X		X	X	X	X	X	X				X	X	X					X		X	X								
Shi *et al.*, 2007 [36]													X			X	X	X	X	X	X	X		X	X		X						
Domingues *et al.*, 2007 [37]	X	X				X	X	X	X	X	X				X	X	X																
Decker *et al.*, 2008 [38]	X	X				X	X	X	X	X	X				X	X	X					X		X	X								
Sánchez-Diz *et al.*, 2008 [39]	X	X				X	X	X	X	X	X				X	X	X					X		X	X								
Padilla-Gutiérrez *et al.*, 2008 [40]	X							X	X	X										X	X	X		X	X								
Goedbloed *et al.*, 2009 [41]	X		X	X		X	X	X	X	X	X				X	X	X					X		X	X								
Ge *et al.*, 2009 [42]	X	X				X	X	X	X	X	X				X	X	X					X		X	X								
Kim *et al.*, 2009 [43]																																	
Ballantyne *et al.*, 2014 [44]	X		X	X		X	X	X	X	X	X				X	X	X					X	X	X	X								X
Robino *et al.*, 2015 [45]																							X										X
Wang *et al.*, 2016 [46]																							X										X

Short Reference	Markers																																
	DYS19	DYS385	DYS385a	DYS385b	DYS388	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS413 (YCAIII)	DYS434	DYS435	DYS437	DYS438	DYS439	DYS443	DYS444	DYS446	DYS447	DYS448	DYS449	DYS456	DYS458	DYS459a/b	DYS460	DYS461	DYS464a/b/c/d	DYS481	DYS505	DYS510	DYS518
Antão-Sousa *et al.*, 2017 [47]	X	X				X	X	X	X	X	X				X	X	X					X		X	X								
Adnan *et al.*, 2018 [48]	X	X				X	X	X	X	X	X				X	X	X					X		X	X								
Bugoye *et al.*, 2018 [49]	X		X	X		X	X	X	X	X	X				X	X	X					X		X	X								
Mertoglu *et al.*, 2018 [50]	X	X				X	X	X	X	X	X				X	X	X					X		X	X								
Yang *et al.*, 2018 [51]	X	X			X	X	X	X	X	X	X				X	X	X		X		X	X	X	X	X								
Wu *et al.*, 2018 [52]	X	X			X	X	X	X	X	X	X				X	X	X	X	X	X	X	X	X	X	X	X	X			X		X	X
Petrovic *et al.*, 2018 [53]	X	X				X	X	X	X	X	X				X	X	X					X		X	X					X			
Zhang *et al.*, 2019 [54]	X	X				X	X	X	X	X	X				X	X	X					X	X	X	X		X			X			X
Ay *et al.*, 2019 [55]																							X										X
Yuan *et al.*, 2019 [56]																							X										X
Ambrosio *et al.*, 2020 [57]	X	X				X	X	X	X	X	X				X	X	X					X		X	X					X			
Lin *et al.*, 2020 [58]	X	X			X	X	X	X	X	X	X				X	X	X		X			X	X	X	X		X			X			X
Fu *et al.*, 2020 [59]	X	X				X	X	X	X	X	X				X	X	X					X		X	X		X			X			

Short Reference	Markers																																
	DYS19	DYS385	DYS385a	DYS385b	DYS388	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS413 (YCAIII)	DYS434	DYS435	DYS437	DYS438	DYS439	DYS443	DYS444	DYS446	DYS447	DYS448	DYS449	DYS456	DYS458	DYS459a/b	DYS460	DYS461	DYS464a/b/c/d	DYS481	DYS505	DYS510	DYS518
Bredemeyer *et al.*, 2020 [60]	X	X				X	X	X	X	X	X				X	X	X					X	X	X	X		X		X	X	X		
Vieira-Silva *et al.*, 2009 [61]	X	X				X	X	X	X	X	X				X	X	X					X		X	X								
Laouina *et al.*, 2013 [63]	X	X				X	X	X	X	X	X				X	X	X					X		X	X								
Soares-Vieira *et al.*, 2008 [62]	X					X	X	X	X	X	X																						

Short Reference	Markers																														
	DYS520	DYS522	DYS526a	DYS526b	DYS527a/b	DYS531	DYS533	DYS547	DYS549	DYS552	DYS557	DYS570	DYS576	DYS587	DYS612	DYS622	DYS626	DYS627	DYS630	DYS635 (C4)	DYS640	DYS643	DYS709	DYS722	GATA A10	GATA H4	YCAII	DYF387S1	DYF399S1	DYF404S1	DYF403S1a	DYF403S1b	DXY156
Lessig & Edelmann, 1998 [21]																																	
Bianchi *et al.*, 1998 [22]																																	
Kayser *et al.*, 2000 [23]																											X						
Dupuy *et al.*, 2001 [64]																																	
Tsai *et al.*, 2002 [25]																																	
Kurihara *et al.*, 2004 [26]																										X							
Dupuy *et al.*, 2004 [24]																																	
Budowle *et al.*, 2005 [27]																																	
Berger *et al.*, 2005 [9]																				X						X							
Gusmão *et al.*, 2005 [29]																				X					X	X							
Ballard *et al.*, 2005 [30]																									X								
Turrina *et al.*, 2006 [31]																				X						X							
Lee *et al.*, 2007 [32]																				X						X							
Hohoff *et al.*, 2007 [33]																											X						X

Short Reference	Markers																															
	DYS520	DYS522	DYS526a	DYS526b	DYS527a/b	DYS531	DYS533	DYS547	DYS549	DYS552	DYS557	DYS570	DYS576	DYS587	DYS612	DYS622	DYS626	DYS627	DYS630	DYS635 (C4)	DYS640	DYS643	DYS709	DYS722	GATA A10	GATA H4	YCAII	DYF387S1	DYF399S1	DYF404S1	DYF403S1a	DYF403S1b	DXY156
Kumagai *et al.*, 2007 [34]																																
Pontes *et al.*, 2007 [35]																				X						X						
Shi *et al.*, 2007 [36]	X					X					X					X			X	X			X		X	X						
Domingues *et al.*, 2007 [37]																																
Decker *et al.*, 2008 [38]																				X						X						
Sánchez-Diz *et al.*, 2008 [39]																				X						X						
Padilla-Gutiérrez *et al.*, 2008 [40]																																
Goedbloed *et al.*, 2009 [41]																				X						X						
Ge *et al.*, 2009 [42]																				X						X						
Kim *et al.*, 2009 [43]		X					X		X			X	X									X										
Ballantyne *et al.*, 2014 [44]			X	X				X					X		X		X	X		X						X		X	X	X	X	X	
Robino *et al.*, 2015 [45]			X	X				X				X	X		X		X	X										X	X	X	X	X	
Wang *et al.*, 2016 [46]			X	X				X				X	X		X		X	X										X	X	X	X	X	

Short Reference	Markers																															
	DYS520	DYS522	DYS526a	DYS526b	DYS527a/b	DYS531	DYS533	DYS547	DYS549	DYS552	DYS557	DYS570	DYS576	DYS587	DYS612	DYS622	DYS626	DYS627	DYS630	DYS635 (C4)	DYS640	DYS643	DYS709	DYS722	GATA A10	GATA H4	YCAII	DYF387S1	DYF399S1	DYF404S1	DYF403S1a	DYF403S1b	DXY156
Antão-Sousa *et al.*, 2017 [47]																				X						X							
Adnan *et al.*, 2018 [48]																				X						X							
Bugoye *et al.*, 2018 [49]																				X						X							
Mertoglu *et al.*, 2018 [50]																				X						X							
Yang *et al.*, 2018 [51]		X			X															X						X							
Wu *et al.*, 2018 [52]	X	X			X	X	X			X	X	X	X	X		X		X	X	X					X	X		X					
Petrovic *et al.*, 2018 [53]							X		X			X	X							X		X				X							
Zhang *et al.*, 2019 [54]							X					X	X					X		X						X		X					
Ay *et al.*, 2019 [55]			X	X			X					X	X		X		X	X										X	X	X			
Yuan *et al.*, 2019 [56]			X	X			X					X	X		X		X	X										X	X	X	X	X	
Ambrosio *et al.*, 2020 [57]							X		X			X	X							X		X				X							
Lin *et al.*, 2020 [58]			X	X	X		X	X	X			X	X		X		X	X		X		X		X		X		X	X	X	X	X	
Fu *et al.*, 2020 [59]							X		X			X	X							X		X				X							
Bredemeyer *et al.*, 2020 [60]		X	X	X			X	X	X			X	X		X		X	X		X		X				X		X	X	X	X	X	

Short Reference	Markers																															
	DYS520	DYS522	DYS526a	DYS526b	DYS527a/b	DYS531	DYS533	DYS547	DYS549	DYS552	DYS557	DYS570	DYS576	DYS587	DYS612	DYS622	DYS626	DYS627	DYS630	DYS635 (C4)	DYS640	DYS643	DYS709	DYS722	GATA A10	GATA H4	YCAII	DYF387S1	DYF399S1	DYF404S1	DYF403S1a	DYF403S1b	DXY156
Vieira-Silva *et al.*, 2009 [61]																				X						X						
Laouina *et al.*, 2013 [63]																				X						X						
Soares-Vieira *et al.*, 2008 [62]																																

3. Discussion and conclusions

Given the rarity of the event itself, the gathering of published STR mutation data and the production of new one is essential to reach relevant insights into mutation rates and corresponding mechanisms. Mutation data obtained from real populations is crucial to identify the genetic profile of different populations, and to obtain more accurate mutation rates. Simple structured Y-STRs (specifically the ones located in the MSY) are the ideal model for this purpose due to their haploid mode of genetic transmission which allows the unambiguous observation of both the paternal and filial alleles involved. Indeed, mutation rates in all systems other than the MSY (X-chromosome or autosomes) may be underestimated due to the occurrence of hidden mutations. The autosomal and X-STRs studies are also of interest since these systems present different modes of genetic transmission, diploid in the former, where recombination occurs both in the maternal and the paternal meioses, and the haplodiploid in the latter, where X recombination occurs in the female meiosis, but no recombination happens in the male meiosis (exempting the PARs). Hence, the autosomes and X chromosome allow the study of the impact of recombination in mutation. The X chromosome, due to its haplodiploid transmission has been of particular interest in these studies, assuming a role neither the autosomes nor its counterpart, the Y chromosome, can fulfil.

Our results support the use of marker-specific mutation rates as these showed to be highly variable between markers (for the Y chromosome from 0.0005 for markers DYS438 and DYS643, to 0.0170 for marker DYS547, and for the X chromosome from 0.0003 for marker DXS7133 to 0.0107 for marker DXS10135). Nevertheless, a high intra-marker variation was also observed, which supports a finer, allele-specific, approach. Longer alleles had higher mutation rates than shorter alleles and presented a tendency to contract (losing repeats). Conversely, shorter alleles presented a tendency to expand (gaining repeats). Mutation rates are frequently more similar between the same allele length from different markers (particularly when considering markers with the same repetitive motif) than between two different alleles in the same marker. Further progression in the state of the art and quantification of allele-specific mutation rates could be drawn from Y-STRs mutation data organized in mutation matrixes, such as the ones we present in this work, in which the specific allele transmission from father to son is made clear.

A positive correlation was observed between the father's age and the occurrence of mutations in Y-STRs, in accordance with previous findings.

Through computer simulations, we have shown that estimating mutation rates using duos for conceals, in general, more mutations than trios. The exception relates to the use of mother-son duos for the X-chromosomal mode of genetic transmission which present the same proportion of hidden mutations as trios when a maternal mutation was simulated. Since the accuracy of the estimates obtained varies so widely depending on the type of familial data considered, the pooling of the data obtained from different pedigrees is incorrect. Moreover, the use of family trios for estimation of STR mutation rates should be favored as opposed to family duos.

The allele frequency distribution of a marker highly influences the proportion of hidden mutations. Therefore, the same marker may show different mutation estimates in distinct populations if the allele frequency distribution is changed, even when it presents the same mutation rate. Since mutations are rare events, and since there was a small variation observed between populations, the pooling of data from different populations (after proper population differentiation analyses) to attain more accurate estimations is however possible. Also, if per marker analyses are performed it is expected that different markers will present distinct levels of confidence for mutation rate estimates.

For autosomal STRs, an inverse correlation between heterozygosity and the frequency of hidden mutations was found in the case of duos.

The X chromosome generally provides more accurate mutation rates estimates than the autosomes (apart from mother-daughter duos), as there are less alleles involved in the transmission, hence, less uncertainty of which parental allele resulted in which filial one. Thus, regarding the X-chromosomal mode of transmission, mother-daughter duos provide the worst estimates, followed by father-daughter duos, mother-son duos, father-mother-daughter trios where a maternal mutation was simulated and, finally, father-mother-daughter trios where a paternal mutation was simulated. Nonetheless, autosomes were found to provide better estimation of the ratio between maternal and paternal mutations than the X chromosome. Hence, biases between paternal and maternal mutation rates differ for autosomal and X-chromosomal modes of transmission, and these should be considered as the level of detection of maternal and paternal mutations is distinct. The mutation rate showed to be higher in the male meiosis.

The high frequency of hidden mutations observed presents a challenge for both single- and multistep mutations. Additionally, multistep mutations are systematically more underestimated than single-step mutations because MIs (Mendelian incompatibilities) are usually interpreted as the change with the least possible number of mutational steps. The framework and the informatics tools proposed allow for the correction of crude mutation rates estimates obtained from the MIA in one-generational familial pedigrees.

Comparing the frequency of single- and multistep mutations in simple-structured Y-STRs, we observed that single-step mutations are indeed more frequent. Nonetheless, the ratio of single- and multistep mutations differs widely between different markers and between groups of markers defined by their repetitive motif structure. At least in some cases, the proportion of single- and multistep mutations is correlated with the structure of the repetitive motif, beyond just its length.

The standardization of mutation data reporting and its availability in every publication is of paramount importance not only for replicability purposes but also to permit the scientific community to reuse and reanalyze it. Hence, we stress the need of the elaboration of recommendations regarding the mode for the publication of results. This assumes crucial relevance as only the standardization in the presentation of results would allow for further advancements in mutation modeling and avoid apparently contradictory conclusions resulting from heterogenous, unreconcilable data.