Towards Genetic Identification with Male-specific Mutations

Arwin Ralf

Author: Arwin Ralf

Cover Design: Anne Morbach

Layout: Arwin Ralf

Printed by: Gildeprint

Towards Genetic Identification with Male-specific Mutations

Richting genetische identificatie met man-specifieke mutaties

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. A.L. Bredenoord

and in accordance with the decision of the Doctorate Board.
The public defence shall be held on

Tuesday 13 December 2022 at 10.30hrs
by

Arwin Ferdinand Ralf
born in Vlissingen, Netherlands.

Erasmus University Rotterdam

Doctoral Committee:

Promotor:	Prof.dr. M.H. Kayser
Other members:	Prof.dr. A.G. Uitterlinden Prof.dr. P. de Knijff Prof.dr. L.M.T. Sijen Prof.dr. L. Roewer Prof.dr. W. Parson Prof.dr. M.E. D'Amato Dr. A.J. Kal
Copromotor:	Dr. M.H.D Larmuseau

CONTENTS

Chapter 1

General introduction and aims of this thesis

Historical forensic science

Forensic sciences have existed for many centuries: in 1773 Carl Wilhelm Scheele devised a chemical method to show the presence of the poisonous arsenic in the body [1], two years later Paul Revere succeeded in identifying a dead body using dental profiling [2]. In the centuries to follow many other forensic field like: forensic ballistics, forensic pathology, forensic psychology and dactyloscopy were pioneered [1]. A significant leap towards the current forensic practice was the discovery of the existence of different blood groups, around the year 1900, by Karl Landsteiner and Paul Uhlenhuth [3] (Figure 1). Determining the blood group of blood that was presumably left at a crime scene by the perpetrator and comparing that to the blood group of a suspect could exclude suspects as being the perpetrator [4]. However, to prove that a suspect committed a crime, blood groups were in no way specific enough as large portions of the population shared the same blood groups. Although the genetic basis of different blood groups was unknown at the time, blood group typing can be considered as the earliest application of forensic genetics.

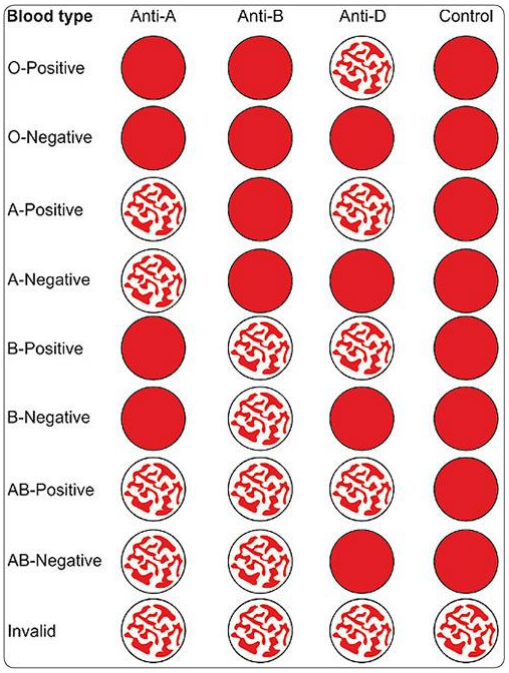


Figure 1: A schematic presentation of the reaction of different blood groups to different antigens (source: Practical Physiology Book).

Scientific revolutions: The discovery of DNA and the birth of 'DNA fingerprinting'

Even before the discovery of different blood groups, in 1869 Friedrich Miescher had first described the existence of a substance in the nuclei of cells which he named nuclein and would later become known as DNA [5]. After its discovery it took over 80 years before James Watson and Francis Crick, in 1953 described the molecular structure [6] with the help of the X-ray diffraction analysis of Rosalind Franklin. In 1958, Matthew Meselson and Franklin Stahl published a study that proved the mechanism of DNA replication as described by James Watson and Francis Crick to be correct [7]. Also in 1958, Crick published the central dogma of molecular biology, stating that DNA sequences can be converted to RNA molecules, which in turn can be translated into protein [8].

The applicability of DNA in forensic sciences to overcome the lack of individual specificity of blood groups was first demonstrated by Alex Jeffreys in 1985 [9], he used variable number tandem repeats (VNTRs) to show individual specific DNA 'fingerprints' (Figure 1). Interestingly, the first application of this new technology in a real forensic case proved the innocence of a man who had previously confessed to having raped and murdered two women [9]. Shortly after, the groundbreaking discovery of Jeffreys was also used to prove that another suspect was the real donor of the latter crime scene sample, consequently that suspect was found guilty and sentenced to life in prison [9].

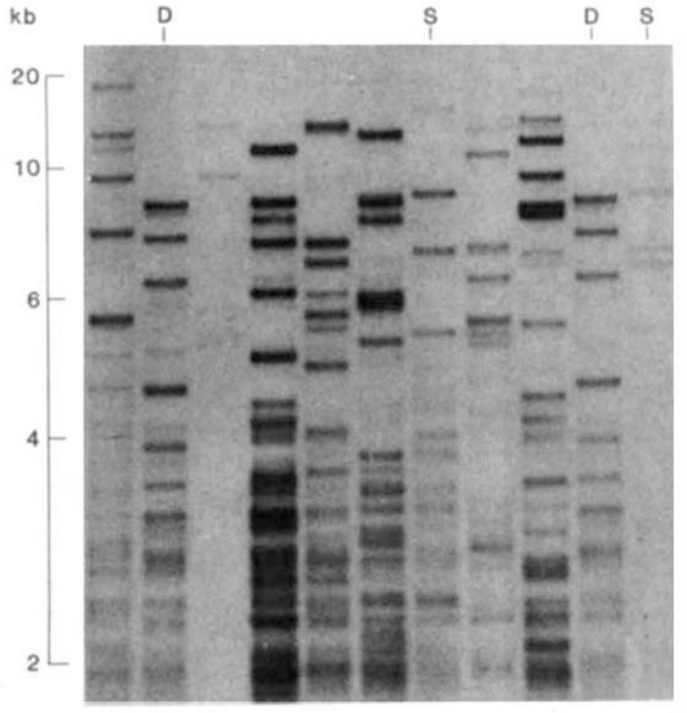


Figure 2: An autoradiograph showing 11 individuals using the restriction fragment length polymorphism based method of analyzing minisatellite repeats (VNTRs) as developed by Alec Jeffreys, the letter D indicates a duplicated individual and the letter S indicates a pair of sisters. The other lanes show the profiles of various other individuals. (Source: A.J. Jeffreys, *et al.*, Individual-specific 'fingerprints' of human DNA, Nature Vol. 316 4 July 1985).

Technology revolutions: PCR and CE

To reach the current state-of-the-art method of forensic DNA profiling, two more technological advances were required. Perhaps the most impactful of all technologies used in the field of molecular genetics was developed by Kary Mullis in 1985: the polymerase chain reaction (PCR) [10]. Because of this method, it became possible to amplify (make many copies of) specific regions of the DNA. Most importantly for forensic genetics, PCR is extremely sensitive, allowing to analyze small traces of biological material [11].

The second noticeable advancement was the invention of the automated DNA sequencer by Lloyd Smith [12], which automated DNA sequencing initially developed by Frederick Sanger in 1977 [13]. The automated DNA sequencer could be used to determine the order of bases (sequence) of an DNA fragment in an automated way by linking it to a computer. The earlier versions of the instruments used large gels to separate the fragments by means of electrophoresis, while later versions used capillaries filled with specific types of polymers for the electrophoretic separation [14]. Capillary electrophoresis (CE), at the moment of writing, is still the most widely used DNA technology in forensic genetics due to its accuracy, cost-effectiveness and the easy at which the results can be interpreted. However, instead of determining the sequence of fragments amplified by PCR, in the forensic field, automated DNA sequencers based on CE are most commonly used for fragment length analysis (FLA) and not for individual base sequencing i.e., to accurately determining the size of PCR-amplified DNA pieces. By using fluorescent labels that are incorporated in the PCR primers and designing the primers in a way so that the regions amplified with the same fluorescent dyes during the same PCR do not overlap in size, it is possible to analyze multiple loci in a single PCR i.e., multiplex PCR [15] and to analyze those targets individually in a single CE run.

STR analysis in forensic genetics

With the rise of PCR, in forensic genetics came the switch from minisatellite repeats or VNTRs in DNA fingerprinting to short tandem repeats (STRs) in forensic DNA profiling as used until today. STRs are relatively short regions (i.e., up to a few hundreds of base pairs) containing a very short motif (i.e., three to six base pairs) that is repeat several times in tandem. The number of repetitions of these motifs varies among individuals; as a result, the total length of the PCR amplicon that includes the STR locus is also variable between individuals. Therefore FLA, in combination with different fluorescent dyes can be used to detect the length of multiple PCR-amplified STRs simultaneously (Figure 3). This approach, which was used from the mid 1990ies [16] until today, has several advantages compared

to the original VNTR-based method: 1) by comparing the length of a given STR to a standardized allelic ladder a repeat number can be assigned to the observed fragment length, as a result STR profiles become standardized and can be exchanged between, or independently produced by any forensic laboratory in the world; 2) because of this standardization the profiles can be stored in databases in the form of a string of numbers, VNTR profiles could not easily be used in that manner; 3) the use of PCR makes the method extremely sensitive, enabling forensic investigators to produce STR profiles from minute amounts of biological material left at a crime scene; 4) the smaller size of the amplified fragments compared to VNTRs allows their detection in degraded DNA; 5) specialized software can be used to automate the interpretation and assignment of repeat numbers to the detected fragments. The statistical power of using STRs for individual identification is astonishing, only about a dozen of polymorphic STRs were sufficient to individualize each human alive, at least statistically speaking. As no one has actually typed the whole human population, empirical evidence for this claim does not exist; nevertheless, two individuals, not being monozygotic twins and yet sharing a full STR profile have not been reported either. Siblings could show, in rare cases, a very high number of shared alleles [17], therefore it is recommended to use slightly more than twelve STRs [18]. Current commercial kits used for individual identification purposes in forensic investigations typically include 16 to 27 autosomal STRs.

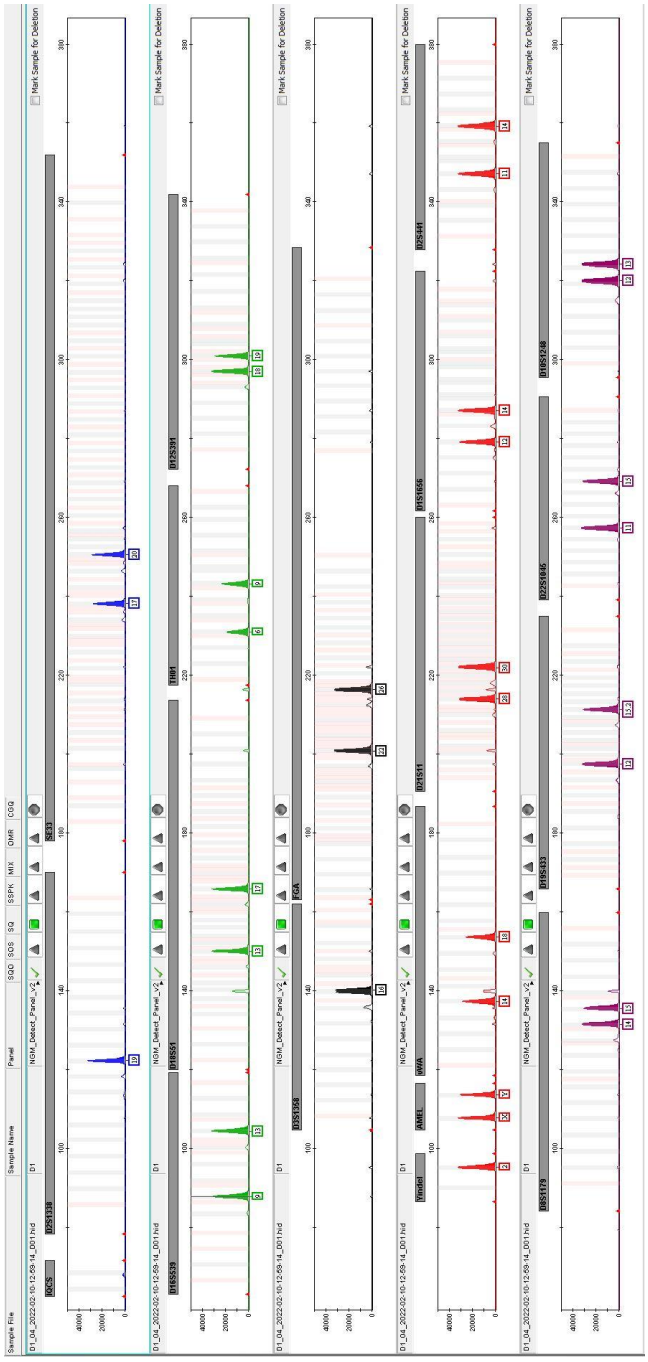


Figure 3: An example of an electropherogram using autosomal STRs amplified by PCR using the six-dye chemistry in the commercial NGM Detect™ PCR Amplification Kit (Thermo Fisher Scientific) including 16 STRs and amelogenin plus a Y-InDel for sex typing, the amplicons were separated and visualized using CE. (Source: The author of this thesis)

The possibility to use national criminal offender DNA databases, which store the STR-profiles of court-convicted offenders together with their personal details such as names and addresses, has revolutionized forensic genetics too. The establishment of such forensic databases and their success in all countries where they have been established is based on empirical data that the repetition rate of criminal behavior is high. For example, an STR profile that is entered in the Dutch criminal offender DNA database has a probability of over 50% to eventually lead to a match with reference sample in the database. The Dutch DNA database contained approximately 350,000 profiles of individuals as of 2021 and on average 85 matches of persons with traces were found weekly [19]. The decision to develop the first STR-based forensic DNA database was made in the United States of America in 1989 by the Technical Working Group on DNA Analysis Methods (TWGDAM), the predecessor of the current SWGDAM (Scientific Working Group on DNA Analysis Methods) [20]. This database would be termed the Combined DNA Index System (CODIS) and was established and is maintained until today by the Federal Bureau of Investigations (FBI). Thirteen autosomal core STR loci were agreed upon in 1997 by a large number of stakeholders: CSF1PO, FGA, TH01, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S359, D18S51, and D21S11 [20]. In 2017, the CODIS set was expanded with another seven loci bringing the total number of STRs to be typed within CODIS twenty [18]. The Dutch criminal offender DNA database was established in 1997 and is hosted by the Netherlands Forensic Institute on behalf of the Dutch Ministry of Justice and Security, and thus not by the police. In other European countries, forensic DNA databases are hosted by police, e.g. the Deutsche DNA-Analyse-Datei (DAD) hosted by the Bundeskriminalamt (BKA) in Germany.

To decide which STRs were suitable to expand the core set of CODIS loci the following criteria were recommended by the CODIS Core Loci Working Group [18]:

- Loci with known associations to medical conditions had to be avoided
- Loci had a low mutation rates (preferably below 3×10^{-3} mpg)
- Loci should not be in linkage disequilibrium (LD) with other included loci
- Loci needed to have a high discriminatory power (preferred prevalence of the genotype of < 0.10%)
- The loci should be, to some degree, be used by the forensic community outside of the US.
- The total number of loci needed to be balanced with their discriminatory power
- Loci needed to comply with the Quality Assurance Standards of the Director of the FBI (i.e., sufficiently validated)

The European counterpart of CODIS is called the European Standard Set (ESS), it originally contained seven Y-STR loci that only partially overlapped with the CODIS loci, and in 2009 five STR loci were added to the ESS [21]. In practice, most commercial STR kits used in forensic DNA analysis target the twelve European core loci and additionally include more STRs (Table 1).

The remarkable capability of autosomal STR loci to differentiate individuals stems from two characteristics. Firstly, being located on different chromosomes, or on distant regions of the same chromosome, their transmission can be considered as independent events. As a result the product rule applies in the statistical interpretation; therefore, the population frequency of a given profile can be estimated by multiplication of the population frequencies of the individually observed alleles at the different STR loci [22]. In contrast, polymorphic variants located on haploid genetic systems, i.e., the non-recombining portion of the Y-chromosome (NRY) and the mitochondrial DNA are not transmitted independently. Therefore, the product rule does not apply to DNA profiles consisting of different Y-chromosomal or mtDNA markers, respectively, and it is typically not possible to perform human identification based on haploid loci; although excluding suspects as contributors to a crime scene trace using haploid markers can be feasible. Secondly, the polymorphic and multi-allelic nature of STRs also provides increased statistical power. Autosomal SNPs could also be used for human identification purposes; however, the number of required SNPs would need to be a few fold higher to achieve a statistical power comparable to that of STRs because SNPs (typically) have only two alleles [23].

Table 1: Different core loci included in CODIS-core, CODIS-20 and ESS and commercial STR typing kits and loci covered by various STR typing kits developed by industry.

				Thermo Fisher Scientific	Thermo Fisher Scientific	Thermo Fisher Scientific	Promega	Qiagen	Thermo Fisher Scientific	Promega	Thermo Fisher Scientific	Qiagen	Thermo Fisher Scientific	Promega	Qiagen	Promega
STR locus	CODIS core	CODIS-20	ESS-12	AmpFLSTR™ Identifiler™	NGM Detect™	AmpFLSTR™ NGM SElect™	PowerPlex® ESX 17	Investigator ESSplex SE QS	AmpFLSTR™ Huaxia™	PowerPlex® 21	GlobalFiler™	Investigator 24plex QS Kit	VeriFiler™ Plus	VersaPlex™ 27PY	Investigator 26plex QS Kit	PowerPlex® Fusion 6C
D18S51	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
D21S11	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
D3S1358	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
D8S1179	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
FGA	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
vWA	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
TH01	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X
D16S539	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X
CSF1PO	X	X		X					X	X	X	X	X	X	X	X
D13S317	X	X		X					X	X	X	X	X	X	X	X
D5S818	X	X		X					X	X	X	X	X	X	X	X
D7S820	X	X		X					X	X	X	X	X	X	X	X
TPOX	X	X		X						X	X	X	X	X	X	X
D12S391		X	X		X	X	X	X	X	X	X	X	X	X	X	X
D1S1656		X	X		X	X	X	X		X	X	X	X	X	X	X
D10S1248		X	X		X	X	X	X			X	X	X	X	X	X
D22S1045		X	X		X	X	X	X			X	X	X	X	X	X
D2S441		X	X		X	X	X	X			X	X	X	X	X	X
D19S433		X		X	X	X	X	X	X	X	X	X	X	X	X	X
D2S1338		X		X	X	X	X	X	X	X	X	X	X	X	X	X
SE33						X	X	X			X	X				X
D6S1043									X	X			X	X	X	
Penta D										X			X	X	X	X
Penta E										X			X	X	X	X

Limitations of autosomal STRs for individual identification

Despite the overwhelming success of using STRs and forensic DNA databases, there are still limitations. The first being obvious: if a perpetrator leaves his/her DNA at the crime scene but is not included in the criminal database because has never been convicted for a crime before, and tactical police investigation does not point to a particular suspect allowing for comparative STR analysis, the perpetrator cannot be identified based on the STR profile generated from the crime scene sample. There is an exception to this rule and that is if a close relative was included in the forensic DNA database; in such cases, a familial search could still point to the perpetrator [24]. However, given the rather limited number of autosomal STRs used, relative identification based on currently used forensic STR profiles typically only allows the identification of first or second degree blood relatives. In many cases, the perpetrator remains unknown, either for ever or at least until his or her STR profile (or that of a close relative) does end up in the database because of having committed another felony.

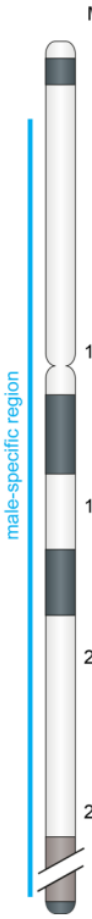
A different challenge for forensic genetics is dealing with mixtures; for example, in sexual assault cases it is common that the DNA of the female victim is present in much larger quantities than the DNA of the male perpetrator. When generating an STR profile from such a mixture the alleles from the perpetrator may be overlapping with those of the victim and hence remain undetected, or they may even not be amplified at all. Such partial STR profiles are less suitable for DNA searches and even in the presence of a suspect may not deliver sufficient evidential value to unequivocally prove that the suspect was indeed the donor of the DNA. A potential solution in these cases could be the use of differential extraction methods [25]; however, this method only works when sperm cells are present and if present, the quantity of those cells needs to be sufficient to generate a good-quality STR profile. Additionally, the success of the approach may vary from case to case, which could depend on the nature of the sample, the specific extraction method that is being used but also on the skill of the analyst performing the extraction [25].

Using Y-STRs in forensic genetics

Another solution in the case of unbalanced male-female mixtures could be the analysis of STRs located on the NRY (Y-STRs). Because biological females don't carry a Y-chromosome, such Y-STR profiles would always be derived from the male perpetrator [26]. The first description of a Y-STR stems from 1992 [27], where the locus DYS19 was described and characterized, and immediately applied in forensic casework [28]. Later, in

1997 a multicenter study suggested the use of a core set of seven Y-STRs for standard Y-haplotyping in forensic casework [29], with the addition of a duplicated locus (DYS385a/b), this set of nine Y-STRs loci was termed the minimal haplotype (MH) [30]. These nine Y-STR loci are still included in the most forensic Y-STR testing kits as of today (Figure 4); although the total number of Y-STRs included in such kits has been constantly expanded throughout time. The reason for further expanding the number of Y-STRs in a kit was the observation that unrelated males shared the same (minimal) haplotypes (i.e., identity-by-state (IBS) without identity-by-descent (IBD) [31, 32]. It was shown that further expanding the number of Y-STRs tested could substantially reduce haplotype sharing between unrelated males [33].

A limitation of using Y-STRs in unbalanced male-female mixtures, or on other sample types confronted with in forensic investigations, lies in the fact that Y-STR profiles typically are not individual-specific, because of the lack of recombination of the region on the Y-chromosome they are located (NRY) and the relatively low mutation rates of standard Y-STRs. As a result, groups of paternally related males typically have the same Y-STR haplotypes; hence, a matching Y-STR profile identifies a paternal lineage with all its individual male members, rather than identifying a single male individual. Paternal lineage identification can be a powerful tool in itself, for example, surnames in most cultures also follow a paternal pattern of inheritance. Therefore, it is proposed that Y-STRs could be used to predict the surnames of perpetrators [34-36]. However, this approach has not yet been shown to be effective in practical casework, because the strong Y-STR / surname correlation only holds for specific surnames. In part, this could be explained by children carrying surnames derived from someone else than their biological father; also, the same surname can have multiple founders in a population [37]. Furthermore, practically applying this approach would require databases covering large numbers of male lineages and Y-STRs haplotype data linked with surname information. Thus, whether this approach will ever become feasible is highly questionable.



Locus	MH (7)	PPY (12)	YFiler (17)	PPY23 (23)	YFiler Plus (27)	Mutation rate / generation
SRY						
DYS393	X	X	X	X	X	0.00107
DYS456			X	X	X	0.00441
AMELY						
DYS570				X	X	0.01140
DYS576				X	X	0.01260
DYS458			X	X	X	0.00617
DYS449					X	0.00944
DYS481				X	X	0.00488
DYS627					X	0.01320
DYS19	X	X	X	X	X	0.00220
centromere						
DYS391	X	X	X	X	X	0.00253
DYS635			X	X	X	0.00421
DYS437		X	X	X	X	0.00132
DYS439		X	X	X	X	0.00546
DYS389I	X	X	X	X	X	0.00272
DYS389II	X	X	X	X	X	0.00433
DYS438		X	X	X	X	0.00351
DYS390	X	X	X	X	X	0.00208
DYS518					X	0.01460
DYS643				X		0.00121
DYS533				X	X	0.00368
YGATAH4			X	X	X	0.00301
DYS385a		X	X	X	X	0.00251
DYS385b		X	X	X	X	
DYS460					X	0.00582
DYS549				X		0.00333
DYS392	X	X	X	X	X	0.00496
DYS448			X	X	X	0.00139
DYF387S1a					X	0.01260
DYF387S1b					X	

Figure 4: An overview of Y-STRs included in different kits, this figure does not include DYS385a/b under the MH set, while these loci are generally considered as part of the MH set making the total number of loci in the MH nine. (Source: Forensic Science Regulator Guidance, FSR-G-227, Issue 1)

Despite having benefits for forensic applications of paternal lineage identification, generally the inability of Y-STR profiles to individualize males, in other words to differentiate all paternal relatives of a man from each other, is regarded as a strong limitation in using Y-STRs as DNA evidence in forensics. In any case, Y-STRs are powerful tools for the exclusion of male suspects from being the trace donor, as a mismatch in a Y-STR profile compared to the crime scene sample proves that the suspect could not have been the donor of that crime scene strain. However, determining the evidential value of a Y-STR haplotype match between a suspect and a crime scene sample, without priors, remains challenging. As all Y-STRs are in LD (i.e., there a dependency between the loci as all loci are transmitted to the next generation together), the product rule cannot be applied to estimate the frequencies of haplotypes in the population. In consequence, in order to weigh the evidential value, an estimation of the prevalence of all loci combined (i.e., haplotype) in the population must be made. Because a Y-STR haplotype consisting of several Y-STRs is way more polymorphic than individual Y-STR loci, such approach could only work if a sufficiently large Y-STR haplotype frequency database was available, which needs to be much larger than the population frequency databased used for retrieving allele frequencies for individuals autosomal STRs.

The need for such a Y-STR haplotype database was already recognized in the early days of Y-STR typing in forensics and in 2001 the first European Y-STR database was launched [38]. This database was the basis for the Y-Chromosome STR Haplotype Reference Database (YHRD), which is the largest Y-STR haplotype population database as of today [39]. However, accurately estimating frequencies of Y-STR haplotypes based on population databases becomes increasingly difficult the more Y-STRs are added to the haplotype (as added to the commercial Y-STR kits), as the expanded haplotypes would typically have a lower prevalence in the population. Therefore, expanding the number of Y-STRs, which is beneficial for improving paternal lineage identification as it decreases IBS and increases IBD would ideally be accompanied by largely expanding the number of haplotypes that such population reference database would include. Moreover, each time a more extensive Y-STR kit becomes available, simply maintaining the number of haplotypes in the database would require retyping all included sample with the new chemistry, which is a costly endeavor [40]. As of today there are 343,932 minimal haplotypes included in YHRD from a total of 1398 populations, while for the most extensive commercially available Y-STR kit (Yfiler Plus), 100,932 haplotypes from 314 population are included (source: https://yhrd.org/pages/resources/stats).

Rapidly mutating Y-STRs

Y-chromosomes do evolve over time by means of mutations and Y-STRs had shown to mutate at an average rate of approximately one mutation every thousand generations [41]. At some point it was hypothesized that there may be Y-STRs that mutate more frequently than others. Such Y-STRs could help overcoming the limitation of haplotype sharing, in particular among paternally related men. To put this hypothesis to the test, Ballantyne *et al.* performed a large study that was published in 2010 [42]. Within this study 186 Y-STRs were analyzed in nearly 2,000 fathers and their sons. The study highlighted a total of 13 Y-STRs that showed an elevated mutation rate of over one mutation every 100 meiotic transfers, i.e., a tenfold increase compared to the estimated average mutation rate of Y-STRs [42]. This class of Y-STRs was termed rapidly mutating Y-STRs (RM Y-STRs) [43]. Subsequently, it was shown in several independent mutation rate studies on father-son pairs from various populations that this set of 13 Y-STRs could differentiated approximately 20-30% of the father-son pairs analyzed [44-53]. Additionally, a study using male pedigrees from Pakistan found that the set of 13 RM Y-STRs could differentiate (i.e., distinguish by at least one mutation) 24% of father-son pairs, 44% of brothers/grandfather-grandson pairs, 55% of uncle-nephew pairs and 61% of first cousins [54]. Although this was a great improvement compared to the state-of-the-art Y-STR genotyping assay at the time [54], a large number of closely paternally related males could still not be differentiated with 13 RM Y-STRs. Furthermore, knowledge on the ability to differentiate males separated by a greater number of generations was scarce.

The Vaatstra-case, a unique case with a unique role for forensic Y-chromosome analysis

In 2012, the Dutch parliament passed an amendment to the Dutch DNA law (Besluit DNA-onderzoek in strafzaken) that allows familial search in the existing criminal offender DNA database and, in case unsuccessful, to carry out large-scale, voluntarily familial searching in specific cases [37]. In the latter approach, a typically large group of males that is selected based on certain criteria, e.g., the place where the crime happened and where they lived at the time of the crime, or their biogeographic ancestry, the men are then asked to voluntarily donate their biological sample (typically a buccal swab) for DNA-based familial search. These men are per definition not regarded as suspects; it is rather hypothesized that they may be to some degree related to the actual but unknown perpetrator. As such familial relationships may be distant, instead of using autosomal

STRs, in this approach the conserved nature of Y-STRs is utilized in Y-STR-based familial search. The general idea for using Y-STR familial search via voluntary DNA dragnets in cases with male perpetrators, which is known e.g. from finding a semen stain on / in a female victim's body and/or from DNA established male sex of the traces sample donor, is that the unknown perpetrator himself does not participate in the voluntary dragnet. Perhaps the most famous case where Y-STR based familial search was used in the Netherlands was the rape and murder of Marianne Vaatstra, a 16 years old girl, in 1999. Even before the mass-screening was allowed by law in 2012, the Y chromosome had played a role in this case. There were suspicions that the perpetrator could have been an inhabitant of an asylum seekers center that was located in the area where the murder had taken place. This suspicion was further fed by the fact that Vaatstra's throat had been slit with a knife, which was regarded as: "something a Frisian would never do" [55]. However, Y-STR analysis, combined with a search in YHRD showed that the perpetrator carried haplotype that was particularly common in males living in Western Europe [37].

After the law had been adapted in 2012 to allow DNA-based familial searching, the Vaatstra-case was the first where a Y-STR based mass-screening was applied. Over 7,500 men that had lived in a radius of 5 km of the crime scene at the time of the murder, in the age group of 16 until 60 years of age, were asked to voluntarily donate their DNA. Hereby, it was communicated that the DNA profiles would not be added to, or compared against the national criminal offenders database and that the profiles and materials would be destroyed after the had been completed. It was also clearly communicated that the DNA dragnet will be conducted based on Y-STRs and that as consequence, a man's participation can result in his non-participating paternal male relatives to become traceable. Nevertheless, an impressive 90% of the invited men decided to donate their DNA.

The power that Y-STR typing could have soon became evident, as the first batch of 81 randomly selected DNA samples already led to two Y-STR haplotype matches with the crime scene stain obtained from the victim's body. This observation confirmed that the perpetrator must have been a local male rather than an asylum seeker. It was noticed, however, that both men had two different surnames, but genealogical research by the investigation team showed that both men shared a common paternal ancestor that had lived nearby in the year 1748 [56]. Based on this result, DNA samples of men that carried these two surnames were prioritized in the Y-STR analysis. On the DNA samples of the males identified by surname and Y-STR haplotyping to belong to this specific paternal lineage, not only Y-STR, but also autosomal STRs analysis was performed to asses if they could match the murderer, or show high similarity which would, in turn, suggest a close familial relationship.

This endeavor went on and more and more Y-STR haplotype matches without autosomal DNA profile matches were unveiled and the investigation team also used more Y-STRs to narrow down the group of close relatives, which are more useful in finding the unknown man than distant ones. At some point in time, and to the surprise of the investigators not expecting the perpetrator to participate, a full autosomal STR match was found between the crime scene DNA and that of one of the voluntary participants. The donor of the DNA found at the crime scene had been identified, in the end because he participated in the voluntary DNA dragnet. He was a local farmer named Jasper S. (Figure 5), who later confessed to having committed the rape and murder of Marianna Vaatstra. After 13 years, this cold case could finally be closed, which is only one of several cold cases in the Netherlands that were and are investigated by means of Y-based familial search. Even if the perpetrator had not voluntarily donated his DNA, it would have been a matter of time before law enforcement would have gotten to him. Ever since those first two Y-STR matches, the net had been closing around him, such can be the power of using Y-STRs to find unknown perpetrators using genealogical searches. Jasper S. was sentenced by the court to 18 years in prison.

Figure 5: A drawing of Jasper S. in court flanked by his criminal defense lawyers (Source: Petra Urban)

Aims of this thesis

The first aim of this thesis was to evaluate if the current set of RM Y-STRs for forensic approaches could be expanded. To achieve this aim a new strategy was employed. Where in the Ballantyne *et al.* study [42] that identified the first set of RM Y-STRs a brute-force approach was employed by characterizing all Y-STR that were available at the time. Here, in **Chapter 2**, we made use of specific characteristics that could be derived from the first set of 13 RM Y-STRs to identify 27 candidate RM Y-STRs that had not previously been characterized. The 27 candidate RM Y-STRs showed molecular characteristics that were similar to the 13 previously identified RM Y-STRs; therefore we hypothesized that they had the potential to display mutability that would classify them as RM Y-STRs, which required empirical validation.

Chapter 3 describes the development and forensic validation of a new method to characterize all previous and newly identified RM Y-STRs efficiently, named 'RMplex'. Before any new method can be applied in forensic casework a developmental validation is required as features like reliability, stability and certainly also sensitive are important to be established as the genetic material found at crime scenes is often scarce and can be of poor quality.

Chapter 4 evaluates RM in an independent set of father-son pairs (and a smaller number of brothers). Independent validation of the characteristics of all Y-STRs included in the method is important to exclude the possibility that the increased mutability of the Y-STRs were specific to the sample set of father-son pairs that were used for the discovery in **Chapter 2**.

While all previous chapters had focused on males of European ancestry, in **Chapter 5** RMplex was evaluated on a set of father-son pairs of East Asian (i.e., Japanese) ancestry. Characterizing the same Y-STRs in different populations is of important as tens of thousands of years of independent evolution of Y-chromosomes may lead to different behavior of specific Y-STRs in various populations.

In **Chapter 6**, we aimed to perform the first comprehensive study on the differentiation of both close and distantly paternally related males with the full set of 30 Y-STRs characterized by increased mutation rates. To do so a large number of male pedigrees were typed with RMplex and Yfiler™ Plus PCR Amplification Kit, the latter being the state-of-the-art Y-STR typing kit developed by industry. Moreover, we aimed to evaluate the potential that RM Y-STRs may have to predict the level of relatedness of two males that carry similar RM Y-STR haplotypes. **Chapter 6,** also includes an elaborate

discussion on how Y-STRs and especially those with high mutation rates could play a much more central role in the field of forensic genetics.

Finally, **Chapter 7** provides a general discussion on the studies that were combined in this thesis, including discussion on remaining limitations of forensic Y-STR analysis that have yet to be overcome and ways to further improve forensic Y-chromosome analysis in the future.

References

1. Hemanth, K., M. Tharmavaram, and G. Pandey, *History of Forensic Science.* Technology in Forensic Science: Sampling, Analysis, Data and Regulations, 2020: p. 1-16.
2. Bruce-Chwatt, R.M., *A brief history of forensic odontology since 1775.* Journal of Forensic and Legal Medicine, 2010. **17**(3): p. 127-130.
3. Patzelt, D., *History of forensic serology and molecular genetics in the sphere of activity of the German Society for Forensic Medicine.* Forensic Science International, 2004. **144**(2-3): p. 185-191.
4. Vogelhut, M.I., *Forensic Applications and Evidential Value of the Blood Group Tests.* University of Detroit Mercy Law Review, 1935. **6**: p. 101.
5. Friedrich, M., *Ueber die chemische Zusammensetzung der Eiterzellen [On the chemical composition of pus cells].* Medicinisch-chemische Untersuchungen, 1871. **4**: p. 441-460.
6. Watson, J.D. and F.H.C. Crick, *Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid.* Nature, 1953. **171**(4356): p. 737-738.
7. Meselson, M. and F.W. Stahl, *The replication of DNA in Escherichia coli.* Proceedings of the National Academy of Sciences, 1958. **44**(7): p. 671-682.
8. Crick, F.H.C. *On protein synthesis*. in *Symposia of the Society for Experimental Biology*. 1958.
9. Jeffreys, A.J., V. Wilson, and S.L. Thein, *Individual-specific 'fingerprints' of human DNA.* Nature, 1985. **316**(6023): p. 76-79.
10. Saiki, R.K., et al., *Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia.* Science, 1985. **230**(4732): p. 1350-1354.
11. Westwood, S.A. and D.J. Werrett, *An evaluation of the polymerase chain reaction method for forensic applications.* Forensic Science International, 1990. **45**(3): p. 201-215.
12. Smith, L.M., et al., *Fluorescence detection in automated DNA sequence analysis.* Nature, 1986. **321**(6071): p. 674-679.
13. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proceedings of the National Academy of Sciences, 1977. **74**(12): p. 5463-5467.
14. Heather, J.M. and B. Chain, *The sequence of sequencers: The history of sequencing DNA.* Genomics, 2016. **107**(1): p. 1-8.
15. Gill, P., R. Sparkes, and C. Kimpton, *Development of guidelines to designate alleles using an STR multiplex system.* Forensic Science International, 1997. **89**(3): p. 185-197.
16. Kimpton, C., et al., *Evaluation of an automated DNA profiling system employing multiplex amplification of four tetrameric STR loci.* International Journal of Legal Medicine, 1994. **106**(6): p. 302-311.
17. Zaken, N., et al., *Can brothers share the same STR profile?* Forensic Science International: Genetics, 2013. **7**(5): p. 494-498.

18. Hares, D.R., *Expanding the CODIS core loci in the United States.* Forensic Science International: Genetics, 2012. **6**(1): p. e52-e54.
19. Meulenbroek, L., *DNA zoekmachine.* 2021.
20. Budowle, B., et al. *CODIS and PCR-based short tandem repeat loci: law enforcement tools*. in *Second European symposium on human identification*. 1998. Promega Corporation, Madison, Wisconsin.
21. Welch, L.A., et al., *European Network of Forensic Science Institutes (ENFSI): evaluation of new commercial STR multiplexes that include the European Standard Set (ESS) of markers.* Forensic Science International: Genetics, 2012. **6**(6): p. 819-826.
22. Kimpton, C.P., et al., *Automated DNA profiling employing multiplex amplification of short tandem repeat loci.* Genome Research, 1993. **3**(1): p. 13-22.
23. Sanchez, J.J., et al., *A multiplex assay with 52 single nucleotide polymorphisms for human identification.* Electrophoresis, 2006. **27**(9): p. 1713-1724.
24. Gershaw, C.J., et al., *Forensic utilization of familial searches in DNA databases.* Forensic Science International: Genetics, 2011. **5**(1): p. 16-20.
25. Klein, S.B. and M.R. Buoncristiani, *Evaluating the efficacy of DNA differential extraction methods for sexual assault evidence.* Forensic Science International: Genetics, 2017. **29**: p. 109-117.
26. Sibille, I., et al., *Y-STR DNA amplification as biological evidence in sexually assaulted female victims with no cytological detection of spermatozoa.* Forensic Science International, 2002. **125**(2-3): p. 212-216.
27. Roewer, L., et al., *Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts.* Human Genetics, 1992. **89**(4): p. 389-394.
28. Roewer, L. and J.T. Epplen, *Rapid and sensitive typing of forensic stains by PCR amplification of polymorphic simple repeat sequences in case work.* Forensic Science International, 1992. **53**(2): p. 163-171.
29. Kayser, M., et al., *Evaluation of Y-chromosomal STRs: a multicenter study.* International journal of legal medicine, 1997. **110**(3): p. 125-133.
30. Roewer, L., et al., *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, 2000. **114**(1): p. 31-43.
31. Pereira, L., M.J. Prata, and A. Amorim. *An evaluation of the proportion of identical Y-STR haplotypes due to recurrent mutation*. in *International Congress Series*. 2003. Elsevier.
32. de Knijff, P., *On the Forensic Use of Y-Chromosome Polymorphisms.* Genes, 2022. **13**(5): p. 898.
33. Larmuseau, M.H.D., et al., *A substantially lower frequency of uninformative matches between 23 versus 17 Y-STR haplotypes in north Western Europe.* Forensic Science International: Genetics, 2014. **11**: p. 214-219.
34. Claerhout, S., et al., *Ysurnames? The patrilineal Y-chromosome and surname correlation for DNA kinship research.* Forensic Science International: Genetics, 2020. **44**: p. 102204.
35. Ochiai, E., et al., *Y chromosome analysis for common surnames in the Japanese male population.* Journal of Human Genetics, 2021. **66**(7): p. 731-738.
36. King, T.E. and M.A. Jobling, *What's in a name? Y chromosomes, surnames and the genetic genealogy revolution.* Trends in Genetics : TIG, 2009. **25**(8): p. 351-360.
37. Kayser, M., *Forensic use of Y-chromosome DNA: a general overview.* Human Genetics, 2017. **136**(5): p. 621-635.
38. Roewer, L., et al., *Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes.* Forensic Science International, 2001. **118**(2-3): p. 106-113.
39. Willuweit, S., L. Roewer, and Y.C.U.G. International Forensic, *Y chromosome haplotype reference database (YHRD): update.* Forensic Science International: Genetics, 2007. **1**(2): p. 83-87.

40. Henry, J., C. Simon, and A. Linacre, *The benefits and limitations of expanded Y-chromosome short tandem repeat (Y-STR) loci.* Forensic Science International: Genetics Supplement Series, 2015. **5**: p. e28-e30.
41. Kayser, M. and A. Sajantila, *Mutations at Y-STR loci: implications for paternity testing and forensic analysis.* Forensic Science International, 2001. **118**(2-3): p. 116-121.
42. Ballantyne, K.N., et al., *Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications.* The American Journal of Human Genetics, 2010. **87**(3): p. 341-353.
43. Ballantyne, K.N., et al., *A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages.* Forensic Science International: Genetics, 2012. **6**(2): p. 208-218.
44. Ballantyne, K.N., et al., *Toward Male Individualization with Rapidly Mutating Y-Chromosomal Short Tandem Repeats.* Human Mutation, 2014. **35**(8): p. 1021-1032.
45. Chen, Y., et al., *Mutation rates of 13 RM Y-STRs in a Han population from Shandong province, China.* Forensic Science International: Genetics Supplement Series, 2017. **6**: p. e346-e348.
46. Javed, F., et al., *Male individualization using 12 rapidly mutating Y-STRs in Araein ethnic group and shared paternal lineage of Pakistani population.* International Journal of Legal Medicine, 2018. **132**(6): p. 1621-1624.
47. Serin, A., et al., *Genetic characterisation of 13 rapidly mutating Y-STR loci in 100 father and son pairs from South and East Turkey.* Annals of Human Biology, 2018. **45**(6-8): p. 506-515.
48. Yuan, L., et al., *Mutation analysis of 13 RM Y-STR loci in Han population from Beijing of China.* International Journal of Legal Medicine, 2019. **133**(1): p. 59-63.
49. Zgonjanin, D., et al., *Mutation rate at 13 rapidly mutating Y-STR loci in the population of Serbia.* Forensic Science International: Genetics Supplement Series, 2017. **6**: p. e377-e379.
50. Zhang, W., et al., *Multiplex assay development and mutation rate analysis for 13 RM Y-STRs in Chinese Han population.* International Journal of Legal Medicine, 2017. **131**(2): p. 345-350.
51. Robino, C., et al., *Development of an Italian RM Y-STR haplotype database: results of the 2013 GEFI collaborative exercise.* Forensic Science International: Genetics, 2015. **15**: p. 56-63.
52. Lang, M., et al., *Comprehensive mutation analysis of 53 Y-STR markers in father-son pairs.* Forensic Science International: Genetics Supplement Series, 2017. **6**: p. e57-e58.
53. Wang, Q., et al., *Rapidly mutating Y-STRs study in Chinese Yi population.* International Journal of Legal Medicine, 2019. **133**(1): p. 45-50.
54. Adnan, A., et al., *Improving empirical evidence on differentiating closely related men with RM Y-STRs: a comprehensive pedigree study from Pakistan.* Forensic Science International: Genetics, 2016. **25**: p. 45-51.
55. Jong, L. and A. M'charek, *The high-profile case as 'fire object': Following the Marianne Vaatstra murder case through the media.* Crime, Media, Culture, 2018. **14**(3): p. 347-363.
56. M'charek, A., *Race and sameness: on the limits of beyond race and the art of staying with the trouble.* Comparative Migration Studies, 2022. **10**(1): p. 1-16.

Chapter 2

Identification and characterization of novel rapidly mutating Y-chromosomal short tandem repeat markers

Arwin Ralf[1], Delano Lubach[1], Nefeli Kousouri[1], Christian Winkler[2], Iris Schulz[2,!], Lutz Roewer[3], Josephine Purps[3], Rüdiger Lessig[4], Pawel Krajewski[5], Rafal Ploski[5], Tadeusz Dobosz[6], Lotte Henke[2], Jürgen Henke[2], Maarten H.D. Larmuseau[7,8], and Manfred Kayser[1]

[1] Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, 3000 CA Rotterdam, the Netherlands
[2] Institut für Blutgruppenforschung LGC GmbH, 50933 Cologne, Germany
[3] Abteilung für Forensische Genetik, Institut für Rechtsmedizin und Forensische Wissenschaften, Charite´-Universitätsmedizin Berlin, 13353 Berlin, Germany
[4] Institut für Rechtsmedizin, Universitätsklinikum Halle, 06112 Halle/Saale, Germany
[5] Department of Medical Genetics and Department of Forensic Medicine, Medical University Warsaw, Warsaw 02-007, Poland
[6] Department of Forensic Medicine, Wroclaw Medical University, Wroc1aw 50-368, Poland
[7]Department of Human Genetics, KU Leuven, Leuven 3000, Belgium
[8]Histories vzw, Mechelen 2800, Belgium

Published in *Human Mutation*, June 24th 2020
Volume 41, Issue 9, Pages 1680-1696
doi: https://doi.org/10.1002/humu.24068

Abstract

Short tandem repeat polymorphisms on the male-specific part of the human Y-chromosome (Y-STRs) are valuable tools in many areas of human genetics. Although their paternal inheritance and moderate mutation rate (~10^{-3} mutations per marker per meiosis) allow detecting paternal relationships, they typically fail to separate paternally related men. Previously, 13 Y-STR markers with untypically high mutation rates (>10^{-2}), rapidly mutating (RM) Y-STRs) were identified and shown to improve male relative differentiation, albeit to a limited degree. By applying a new *in silico* search approach, we identified 27 novel RM Y-STR candidates. Genotyping them in 1,616 DNA-confirmed father-son pairs for mutation rate estimation empirically highlighted 12 novel RM Y-STRs, for which we estimated the capacity to differentiate males related by 1, 2, and 3 meioses at 27%, 47%, and 61%, respectively, while for all 25 currently known RM Y-STRs it was 44%, 69%, and 83%. Of the 647 Y-STR mutations we observed in total, almost all were single repeat changes, repeat gains and losses were well balanced; allele length and fathers' age were positively correlated with mutation rate. We expect these new RM Y-STRs, together with the previously known ones, to significantly improving male relative differentiation in future human genetic applications.

Introduction

Short tandem repeat (STR) analysis has grown over the last 25 years to become and remain the gold standard for human individual identification purposes in forensic genetics [1, 2], while they are also used in other human genetic areas. Besides autosomal STRs, the human genome of male individuals also contains hundreds of STRs located on the male-specific portion of the human Y-chromosome (Y-STRs). Such male-specific Y-STR markers have become increasingly popular in various areas of human genetics such as in forensic genetics [3], genetic genealogy [4], anthropological genetics and human population history research [5].

In forensic genetics, Y-STRs are especially useful for solving sexual assault cases with DNA mixtures typically containing an excess of DNA from the female victim's epithelial cells compared to DNA of the male perpetrator's sperm cells [6]. Based on such imbalanced male-female DNA mixtures, it often is practically impossible to identify the male contributor based on autosomal STR profiling, even after differential lysis leading to enrichment of sperm DNA was applied [7, 8]. In contrast, a Y-STR profile (haplotype) of the male contributor can typically be obtained from such mixed material, which allows determining the paternal lineage to which the male crime scene trace donor belongs [3]. Because of the lack of recombination and the relatively low mutation rate (~10^{-3} mutations per marker per meiosis) of the Y-STRs typically used in forensic Y-chromosome analysis, a Y-STR haplotype highlights the male perpetrator together with many of his paternally related male relatives. This allows particular forensic Y-STR applications of genetic identification such as familial searching [3], forensic genealogy [9], or surname prediction [10]. In general, however, forensic DNA analysis seeks individual identification.

Male relative differentiation using Y-chromosome markers is achievable by using Y-STRs with a high mutation rate. However, for almost two decades of Y-STR research and applications, only Y-STRs with moderate mutation rates in the order of 10^{-3} mutations per marker per meiosis were known. This situation changed in 2010 with the publication of a large empirical Y-STR mutation rate study analyzing 186 Y-STRs in nearly 2000 DNA-confirmed father-son pairs, which highlighted 13 Y-STR markers with mutation rates >10^{-2} mutations per marker per meiosis termed Rapidly Mutating (RM) Y-STRs [11]. Followed by the first empirical demonstrations of their suitability for male relative differentiation [12, 13], many subsequent studies provided increasing evidence on the value of RM Y-STRs for differentiating related, including closely related, and also unrelated men [14-24]. In genetic genealogy too, RM Y-STRs are advantageous as they provide improved

differentiation of unrelated individuals [13] and they allow distinguishing closely related from more distantly related males by taking the number of observed mutations into account [25].

However, the relatively small number of 13 previously identified RM Y-STRs provides limitations for male relative differentiation, particularly regarding closely related men, which limits applications in forensic genetics and genetic genealogy [26]. Empirical studies based on hundreds of male relative pairs showed that these 13 RM Y-STRs allow separation of males related by one, two, three, and four meioses with 27%, 46%, 54%, and 62%, respectively [17], which demonstrates room for improvement. This shortcoming in the male relative differentiation rates of the previously identified RM Y-STRs motivated our search for additional RM Y-STRs, which - if identifiable - are expected to further improve male relative differentiation, particularly of closely related men.

There are different approaches to estimate mutation rates of Y-STRs serving as prerequisite for classifying Y-STRs as RM Y-STRs (i.e. $\mu > 10^{-2}$ mutations per marker per meiosis). One approach is the use of DNA-confirmed father-son pairs [11, 27]; however, for revealing reliable mutation rate estimates with this approach, the number of analyzed father-son pairs needs to be large. Alternatively, a high-resolution Y-SNP based phylogeny in a population-based approach [28], or deep-rooted male pedigrees [23, 29] could be used to estimate mutation rates of Y-STRs. The latter two approaches require less individuals to be genotyped to cover the same number of generations compared to a father-son based approach . This is especially beneficial for estimating the mutation rate of Y-STRs with moderate to low mutation rates (i.e. $\mu \sim 10^{-3}$ and less) [28]. For such Y-STR markers the father-son based approach requires thousands, or even tens of thousands of pairs to obtain reliable mutation rate estimates. However, for RM Y-STRs with mutation rates $>10^{-2}$, the number of father-son pairs required to achieve reliable mutation rate can be lower, i.e., analyzing one thousand father-son pairs expects to find at least 10 RM Y-STR mutations. Moreover, population-based approaches and to some extent deep-rooted pedigree analysis, rely on assumptions regarding the number of generations from the tested individuals to the most recent common ancestor (MRCA), which can lead to inaccurate estimations of the mutation rates [28, 30]. Another disadvantage of both of these approaches is the potential presence of parallel mutations, hidden mutations and multi-step mutations, which all could lead to increased error in the mutation rate estimates obtained [31]. Therefore, particularly for RM Y-STRs, direct observation in father-son pairs, provided a sufficiently large number of pairs being available for analysis, represents the preferred approach for establishing mutation rates. Moreover, only this approach allows characterizing the direction of the repeat mutations (repeat gain versus

repeat loss) and quantifying the step-wise nature of the repeat mutations (single step versus multi-step) unambiguously.

Since our previous Y-STR mutation study [11] already included most Y-STRs known at the time, but only identified 13 RM Y-STRs, in the present study aiming to find additional RM Y-STRs, we had to use a different approach. First, we developed an *in silico* method that can identify (Y-)STRs with increased mutation rates. Next, we applied this *in silico* search method to the Y-chromosome reference sequence (GRCh38) to identify novel RM Y-STR candidate markers. Then, we genotyped the identified candidate RM Y-STR markers in over 1,600 DNA-confirmed father-son pairs to establish their mutation rates, which empirically identified RM Y-STRs out of the *in silico* highlighted candidate markers. We also provide a first expectation on the male relative differentiation capacity these novel RM Y-STRs provide and compared them with the previously known RM Y-STRs. Lastly, by taking advantage of the large number of Y-STR mutations we observed among the large number of father-son pairs, we analyzed the obtained mutation data regarding the impact of allele length, father's age at time of conception, and repeat motif sequence composition on Y-STR mutation rates to gain further insights into the mutability of Y-STRs in general.

Materials & Method

Editorial Policies and Ethical Considerations

The use of all completely anonymized DNA samples for the purpose of this study was in agreement with the institutional regulations and was under informed consent.

Candidate RM Y-STR marker ascertainment

We identified candidate RM Y-STR markers (cRM Y-STRs), by scanning the entire Y-chromosome reference sequence. In particular, we first built a catalogue containing all Y-STRs present in the latest assembly of the human genome (GRCh38), by using the publically available software Tandem repeats finder [32]. The following parameters were set in the software: Match = 2, Mismatch = 100, Delta = 100, PM = 80, PI = 10, Minscore = 12, MaxPeriod = 5. These settings resulted in a catalogue containing only uninterrupted (perfect) STRs with a maximum repetitive motif size of five base pairs. For the purpose of this study, only STRs located on the Y-chromosome were considered. From the resulting Y-STR catalogue we discarded all repeats with a motif size below three, as such markers

suffer from too much stutter [33]. Y-STRs located in pseudoautosomal regions (PAR) were also excluded, because such regions do not contain male-specific loci [34, 35]. Y-STR markers of which the mutation rates were comprehensively estimated in a previous study (Ballantyne et al. 2010) were excluded too. On the resulting cleaned catalogue, we used a top-down approach where we first attempted to design primers for the cRM Y-STRs with the highest number of repeats. If a single uninterrupted repeat stretch had another (preferably long) repeat in close proximity, i.e. <200 base pairs, we attempted to design primers in such a way that both repeat stretches would be included. We also enriched the set for multi-copy loci by favoring these loci over single-copy loci with the same repeat length in the reference genome when considering Y-STR markers for primer design.

To predict, which STR locus is prone to expressing high mutability, we developed a workflow that can assign a mutability prediction score to any STR sequence. For calculating this score, we used—in a locus specific way—four molecular features that had previously shown to impact on (Y-)STR mutability [11, 28, 36-41] : i) the length (i.e. number of repeats) of the uninterrupted repeat stretches, ii) the number of repeat stretches in a sequence, iii) the marker being a single-copy, or a multi-copy marker, iv) the size (i.e. number of base pairs) of the repeat motif. Of these features, the length of the uninterrupted repeat stretches was previously shown to be the most important factor increasing (Y-)STR mutation rates [11, 36-41].

In order to assign the mutability prediction score to a given Y-STR marker, first the sequence was converted to an "STR structure sequence", which counts the repeats stretches with more than four repetitive units in the following systematic way. For each repetitive sequence belonging to the same motif sequence family, a single repeat nomenclature was applied. For instance, $[AAAG]_n$, $[AAGA]_n$, $[AGAA]_n$, and $[GAAA]_n$ as well as their complementary sequences $[TTTC]_n$, $[TTCT]_n$, $[TCTT]_n$, and $[CTTT]_n$ were all counted as one motif sequence family $[AAAG]_n$. Examples using two previously published Rapidly Mutating Y-STRs are shown in Figure 1. Next, the converted STR structure sequences were used as input for our algorithm to assign the mutability prediction score. In the case of multi-copy markers, the sequences of the different copies were concatenated into one sequence representing all copies together. Total repeat length has previously shown exponential correlation with Y-STR mutability [11, 37, 38, 41], therefore an exponential function was derived empirically from the Y-STRs and mutation rates described previously [11]. The score assigned to each uninterrupted repeat stretch can be expressed as $e^{(0.15 \times \text{number of repeat units})}$; if multiple uninterrupted repeats were present, the scores of the individual uninterrupted repeats were summed up. For example, the previously identified RM Y-STR DYS627 [11] contains two repeat stretches, one of six and one of eighteen

repeats in the Y-chromosome reference sequence (GRCh38) (Figure 1); thus, the score assigned to this RM Y-STR is $e^{0.9} + e^{2.7} = 2.46 + 14.88 = 17.34$. The other previously identified RM Y-STR used as an example in Figure 1, DYS526b, has three repeat stretches and received a score of $e^{2.1} + e^{1.35} + e^{1.95} = 19.12$. Lastly, tetranucleotide repeats were previously found to be more mutable than other motifs, i.e. trinucleotide, or pentanucleotide repeats, when considering similar numbers of repeat units [11, 38]. Therefore, if the repeat motif –predominantly- belonged to any other motif size class, the final score was adjusted by dividing it by 2 (mononucleotide and dinucleotide repeats were not considered in this study).

Previously, information about Y-STRs, i.e. nomenclature and genomic locations etc. were stored in the Human Genome Database (GDB), which, however, is no longer available. In order to verify whether the cRM Y-STRs were already described previously, we searched for the genomic locations of the cRM Y-STRs in "ISOGG YBrowse" (https://ybrowse.org). Table S1 shows the nomenclature for the markers that were already described, although no comprehensive mutation rate estimates were available for these markers. Additionally, for the cRM Y-STRs that were not found in the browser, or those that only partially overlapped with known Y-STRs, we proposed new names (Table S1). We assigned DYS-numbers to single-copy markers and DYF-numbers to multi-copy markers. We used numbers larger than one thousand since such numbers had not yet been used to describe Y-STRs.

DYS627

CTAGGTGACAGCGCAGGATTCCATCTAAAAACAAAAATCAAAGAGAGACACAGGGAGAAAGAAAGAAAGGAAAAAAGAAAGAA
AGAGAGAGAGAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAA
GAAAGAAAAAGAGAAAGAGAGAGGGAGGGAGGAGGGAGGGAGGAAGTAAAGAAGGTAAGAAGGAAGGATGGAAGTAAGGAAG
GAAAGAAGGAAGGAAGGAACGAAGGAGAGAATGAAAGGCATAAAACCAAAGGGAAATGAATAAAAATAAAAACTTGCCATTTG
CTCATTATCC

[N]83[AG]6[AAAG]18[N]170

DYS526b

TCAGCAAACACCTTAGCTAATCATGATTCTTTACCTTCTGGTGAACTGATCCAAACCTTTACTTATTTTCTTCCTTTTTTCCTCTCTCC
TTCCCTCCCTCCCTCTTTCTTCCTTTCTTTCTTC**CTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCCC**
TTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCTTTCCTTCTTTCCTTCATTTCTTCTTTCTTTCTCTTTACCTTCTGGGAAACTG
ATCCAAACCTTTACTTACTTTCTATTTTTCTCTATTTTTCTTTCTCTATTTTTTTCTCT**CCTTCCTTCCTTCCTTCCTTCCTTCCTTCC**
TTCCTTCCTTCCTTCCTTCCTTCCTTCTCTTTACCTTCTGGCGAAGTAACCC

[N]26[AAGG]14[N]113[AAGG]9[AAAG]13[N]123

Figure 1: Examples of the conversion of full STR marker sequence to STR structure sequence for two previously identified RM Y-STRs, DYS627 and DYS526b, as part of the newly developed in silico approach used to find novel RM Y-STRs, for illustrative reasons. Note that for DYS526b the reverse complementary sequence was used to meet the "single motif requirement" (see materials and methods for explanation).

Primer design, multiplex development, and genotyping

The cRM Y-STRs identified with the *in silico* approach were followed-up by genetic testing in father-son pairs to establish their mutation rates empirically and thus demonstrate their RM Y-STR status. For this, PCR primer design was performed using Bisearch [42] to estimate the melting temperature of the primers. Bisearch was also used to perform *in silico* PCR in which only Y-chromosome specific *in silico* amplicons were allowed. Lastly, Bisearch was used to ensure that individual primers were reasonably specific, i.e. did not bind to many hundreds, or thousands of locations across the human genome. All primer pairs that were designed were first tested by performing singleplex PCRs on both male and female DNA samples to ensure male-specific amplification; for this, the PCR products were visualized on agarose gels. In cases where amplification in female samples was observed, PCR primers were redesigned. If also redesigning the primers did not lead to male-specific amplification, capillary electrophoresis (CE) was used to check if the unspecific amplicons overlapped with, or were in close proximity (<20 base pairs) to, any of the known alleles from Y-STR loci within the same fluorescent dye channel. If this was not the case, the marker remained in the study; if this was the case, the marker was excluded from further analyses. Of the 38 cRM Y-STRs considered for primer design, 11 were excluded due to unspecific amplification overlapping with male-specific products despite our attempts.

In total, we successfully designed PCR primers for 27 cRM Y-STRs; those 27 markers were divided between six multiplex PCR assays to allow more efficient (compared

to singleplex PCR) genotyping of the large number of DNA samples from fathers and their sons we considered in this study. Autodimer software [43] was used to ensure the primer combinations had minimal primer interactions. Oligonucleotides targeting the 27 cRM Y-STRs were purchased with 5' labeling of the forward primer using either 6-Fam, Joe, or TAMRA (Metabion International AG). Primer sequences and additional information, i.e. primer sequences and mutability prediction scores, of the cRM Y-STRs can be found in Supplementary Table S1. Each multiplex was optimized using five high-quality human male DNA samples, one high-quality female human DNA sample and two negative control samples. PCR reactions were performed in 10 µL volumes, containing 5 µL of QIAGEN Multiplex PCR Master Mix (QIAGEN N.V.), oligo nucleotides at varying concentrations ranging from 0.1 to 1 µM, and 1 µL of template DNA. While concentrations of template DNA added with 1 µL to the PCR reaction varied, peak height inspections in the electropherograms demonstrated that genotype data for all samples and markers analyzed were reliably obtainable. The PCR reactions were performed on GeneAmp PCR System 9700 (Thermo Fisher Scientific Inc.) using both 96-well and 384-well dual blocks. Every multiplex reaction was amplified with the same PCR protocol: 94 °C for 10 min, 10 cycles of 94 °C for 30 s, 65-1 °C every cycle for 60 s and 72 °C for 60 s, followed by 25 cycles of 94 °C for 30 s, 50 °C for 30 s and 72 °C for 60 s with a final extension step of 60 °C for 45 min. After amplification, 1 µL of the PCR product was mixed with 9 µL of Hi-Di formamide (Thermo Fisher Scientific Inc.) and with 0.3 µL of ILS600 size standard (Promega Corporation). This mixture was incubated at 95˚C for 3 minutes and rapidly cooled on ice for 5 minutes. Capillary electrophoresis was performed on an ABI3130XL Genetic Analyzer (Thermo Fisher Scientific Inc.) using sixteen 36 cm capillaries and POP-7 Polymer (Thermo Fisher Scientific Inc.). The Any4Dye spectral calibration matrix (Promega Corporation) was installed which allowed for accurate separation of signal from the different fluorescent labels. The resulting electropherograms were analyzed using Genemapper software version 4.0 (Thermo Fisher Scientific Inc.).

The newly developed multiplex systems to analyze the 27 cRM Y-STR were then used to genotype 3,232 DNA samples which were derived from sample donors of German and Polish European descent, representing a total of 1,616 DNA-confirmed father-son pairs. These samples are a subset of the father-son pairs used in our previous comprehensive Y-STR mutation rate study [11], excluding samples with DNA shortage, or incomplete amplification of all markers of the father's and/or the son's DNA of a given pair. The true biological father-son relationship was previously established by means of autosomal DNA-analysis; more detailed information about the samples can be found in the initial publication [11]. Data interpretation was performed independently by two

research technicians and conflicting results were resolved by a third trained specialist. If an allelic difference had been observed within a given father-son pair at any cRM Y-STR tested, the result was confirmed by independent genotyping of both father and son to confirm the allelic difference before concluding that the allelic difference reflected a mutation. In the case of multi-copy markers it was decided that peak height ratio differences would not be interpreted as mutations, e.g., a hypothetical multi-copy marker could mutate from 15-15-16 to 15-16-16, resulting in an increased peak height for allele 16 and a decreased peak height for allele 15 in the son. However, there are other factors that can influence the peak height ratios, e.g., preferential amplification of one or more alleles as a result of primer binding site mutations, or a stochastic amplification bias as a result of a low amount of input DNA. Therefore we preferred a conservative approach and ignored such peak height differences in the mutation analysis of multi-copy markers i.e., call both the father and son as 15-16 in the example given above.

Mutational data analysis

Validation of mutability prediction score

In order to validate whether the mutability prediction score was a suitable predictor for Y-STR mutation rate, 185 Y-STRs from our previous mutation rate study [11] were grouped, according to their mutation rates, as follows: slowly mutating Y-STRs (SM Y-STRs): n=82, with mutation rates $<10^{-3}$ mutations per marker per meiosis (in the following used without the unit of measure); moderately mutating Y-STRs: n=70 with mutation rates $\geq 10^{-3}$ and $<5.0\text{x}10^{-3}$ (MM Y-STRs); fastly mutating: n=19 mutation rates $\geq 5.0\text{x}10^{-3}$ and $<10^{-2}$ (FM Y-STRs); and rapidly mutating Y-STRs: n=14 mutation rates $\geq 10^{-2}$ (RM Y-STRs). Note that the A and B parts of the multi-copy RM Y-STR marker DYF403S1 were considered separately in this analysis, DYF403S1b has a size range that is clearly distinguishable from the allele range of DYF403S1a. Therefore these a and b parts were analyzed separately and for both parts the mutation rates were estimated separately. The statistical significance of the differences in the mean mutability prediction scores between these four groups were tested using pairwise Wilcoxon rank sum test and with Bonferroni p-value adjustments for multiple testing in RStudio (https://rstudio.com).

Mutation rate estimation

Mutation rates were calculated in a locus-specific manner using the frequentist approach i.e., dividing the total number of observed mutations for a Y-STR marker by the total number of father-son pairs tested for a Y-STR marker; the mutation rate is therefore expresses as the number of mutations per marker per meiosis. Estimating the mutation rates of individual repeat stretches within complex STR loci, or estimating the mutation rates of individual copies in multi-copy loci was not possible with genotyping methodology that was used. The 95% confidence intervals of the mutation rates were calculated with the Clopper-Pearson (exact) method using a binomial distribution in RStudio, using the "exactci" package.

Differentiation capacity estimation

To provide a first expectation to what degree the identified novel RM Y-STRs will improve differentiating male relatives, the theoretical differentiation capacities (r_d) were calculated for different Y-STR marker sets (from i = 1 to n; with n being equal to the number of Y-STR markers in each set) based on estimated mutation rates (r_m) for different numbers of separating meioses (m) using the formula:

$$r_d = 1 - \prod_{i=1}^{n} (1 - r_m)^m$$

Testing mutation effects of allele length

To test the effect of fathers' allele lengths on Y-STR mutation rate and the direction of mutations, a categorical approach was used. Categories were defined within each marker using the tertiles, where the low range was defined as alleles with the length equal to, or lower than the first tertile allele, the medium range consisted of the alleles greater than the first tertile and smaller then, or equal to the second tertile, the high range was defined as all alleles greater than the second tertile. The number of alleles and the mutations within these three categories were summed up across all markers. To statistically test if allele length had a significant impact on the mutability, the allelic mutation rates, i.e., the number of mutations per allele per meiosis, between the three categories were compared using pairwise comparison of proportions, combined with Bonferroni p-value adjustments in RStudio. To statistically test if the allele length has a significant impact on the direction of the mutations, the proportions of expansions and contractions within the three categories were calculated using exact binomial testing in RStudio.

Testing mutation effect of father's age at the time of son's conception

To test if there was a significant effect of the father's age at the time of conception on the Y-STR mutability, all fathers of which age information was available (N=1,500) were grouped in four age categories by using the quartiles. Group 1 consisted of 432 fathers with ages below 24 at the time of conception; group 2 ranged from age 24 to 29 and contained 378 individuals; group 3 ranged from age 30 to 36 with 324 individuals; and group 4 contained fathers that had reached age 37 and beyond at the time of conception and contained 366 individuals. To test if there were statistically significant differences between these age groups in the number of mutations that occurred, we used pairwise comparisons of the mean number of mutations per individual in each age groups using Wilcoxon rank sum test and with Bonferroni p-value adjustments in RStudio.

Testing mutation effect of repeat motif sequence

To test for the influence of the repeat motif sequence on Y-STR mutation rates, eight commonly found motif sequences families, specifically: AAG, AGG, AAT, AAC, AAAG, AAGG, AGAT, and AAAT, were compared between RM Y-STRs and non-RM Y-STRs. The non-RM Y-STRs were ascertained from a previous study [11], while for the RM Y-STRs, the 13 markers identified in the same previous study were combined with the novel RM Y-STRs identified in the present study. Two-tailed Fisher's exact test, in RStudio, was used to test for significant differences in motif sequence composition between the RM and non-RM Y-STRs.

Results & Discussion

Candidate RM Y-STR marker ascertainment

Estimating to what degree the developed and applied mutability prediction scores actually correlate with mutability, we first performed a linear regression analysis of the mutability prediction scores with the empirically derived mutation rate estimates for 185 Y-STR markers from our previous mutation study including the 13 known RM Y-STRs [11]. A statistically significant positive correlation was observed with an R^2 of 0.53 (p-value < $2.2x10^{-16}$). However, a limitation of the used dataset is that it contains many markers (51% of total Y-STRs analyzed) with either just a single, or no mutation observed in the nearly 2000 father-son pairs analyzed in the previous study. This makes the mutation rates estimated for such markers less reliable [28] with an expected impact on our correlation analysis. To gain more insights into the effect of mutation rate uncertainty on our mutability score correlation analysis, we additionally applied a categorical approach on the same dataset to visualize the differences in mutability prediction scores between Y-STR markers using four marker groups defined by their mutation rates: slowly mutating (SM Y-STRs), moderately mutating (MM Y-STRs), fast mutating (FM Y-STRs) and rapidly mutating Y-STRs (RM Y-STRs) (for mutation rate definitions of these groups see method section 2.4). SM Y-STRs showed significant p-values (Wilcoxon rank sum test) compared to all other three groups MM Y-STRs, FM Y-STRs, and RM Y-STRs (p-values of $1.7x10^{-7}$, $3.6x10^{-7}$, $1.7x10^{-8}$, respectively). MM Y-STRs showed significant p-values compared to FM Y-STRs and RM Y-STRs (p-values of 0.0092 and $7.2x10^{-8}$, respectively). Comparing FM Y-STRs to RM Y-STRs resulted in a significant p-value of 0.0076.

As evident from Figure 2, a mutability prediction score of >15 provides reasonably good indication for RM Y-STRs, although finding some markers with slightly

lower mutating rates can also be expected when using such mutability score threshold. Importantly, for the 27 cRM Y-STRs highlighted in our *in silico* analysis and included in the multiplex genotyping, the mean mutability score was 33, ranging from 7 to 123 across markers (Table S1). Moreover, based solely on the length of the longest repeat stretch, 7 of the 13 previously described RM Y-STRs [11] were found among the top candidates (before taking multiple repeat stretches and multi-copy status into account), which demonstrates the suitability of our *in silico* approach, including the use of our mutability score, to find RM Y-STR markers, and provides promises that we can find new RM Y-STRs with our *in silico* approach.

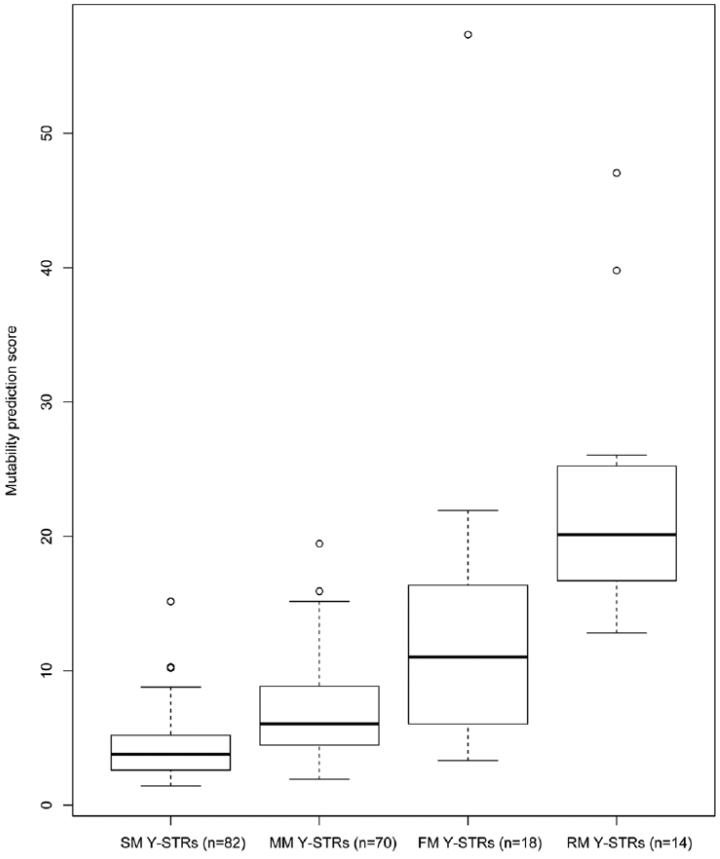


Figure 2: Boxplots showing the distributions of the newly developed mutability prediction scores among four groups of Y-STR markers as defined by mutation rate: i) slowly mutating (SM) Y-STRs (mutation rate <10-3.), ii) moderately mutating (MM) Y-STRs (mutation rate ≥10-3, < 5x10-3), iii) fast mutating (FM) Y-STRs (mutation rate ≥5x10-3, <10-2), and iv) rapidly mutating (RM) Y-STRs (mutation rate ≥10-2) based on Y-STRs and their mutation rate estimates from Ballantyne et al., 2010.

Mutation analysis

Genotyping the 27 cRM Y-STR markers in 1,616 DNA-confirmed father-son pairs revealed a total of 647 repeat mutations across all markers and pairs. The mean number of mutations per marker was 24, ranging from two to 84 across markers. A positive correlation of the empirically derived marker specific mutation rate with the mutability prediction score was observed (R^2 of 0.66, $p=3.8 \times 10^{-7}$). Of the 647 mutations, 318 (49%) were repeat expansions and 322 (50%) were contractions, demonstrating a nearly equal ratio. This finding differs slightly from that of our previous study based on 186 Y-STRs selected independent of mutation rate expectation, where of the 787 mutations observed in total, slightly more repeat contraction (423; 54%) than repeat expansions (364; 46%) were found [11]. For seven mutations in our present study, the direction could not be unambiguously assigned due to the multi-copy status of the involved markers, explaining the missing percent. For instance, observing within a father-son pair the genotype combinations 15-16-17 and 15-17 could mean a mutational repeat loss from 16 to 15 or a repeat gain from 16 to 17, or alternatively a deletion of the locus copy with allele 16. Although the repeat gains versus losses were equal across all cRM Y-STR markers, four markers showed large differences in the directionality of the mutations. In DYS1003 and DYS1013 repeat contractions were dominant with 76% and 75%, respectively (p-values of 0.012 and 0.077, respectively), while in DYS1006 and DYS1017 it were predominantly repeat expansions with 78% and 77%, respectively (p-values 0.180 and 0.092, respectively). However, these differences only led to a significant p-value in one single marker (i.e. DYS1003), which may be explained by the lower number of observed mutations in the remaining three markers. Future research will have to show if these observations can be confirmed with additional mutations found by analyzing additional father-son pairs.

For the analysis of the step-wise nature of the mutations, two markers, namely DYF1000 and DYS1010, were excluded from this analysis, since the sequences contain both trinucleotide repeats combined with a hexanucleotide repeat, and tetranucleotide repeats combined with a dinucleotide repeat, respectively. Hence, in the case of DYF1000, finding a mutation with a six base pair difference could be explained as either a single-step mutation of the hexanucleotide repeat, or as a two-step mutation of the trinucleotide repeat (or even as two single-step mutation at different trinucleotide repeat stretches). Similarly, in DYS1010 a four base pairs difference in a father-son pair could be explained as either a single-step tetranucleotide mutation, or a two-step dinucleotide mutation. The vast majority of the 563 mutations observed in the remaining 25 cRM Y-STRs were single-

step repeat mutations (544, 97%, Table 1), which agrees well with the results from our previous study with 96% single-step mutations [11]. In the present study, only 3% of the observed mutations were two-step mutations and less than 1% were three-step mutations (Table 1). Notably, our present dataset contained two individuals (both were sons) that appear to carry a large deletion in their Y-chromosomes, resulting in a large number of null-alleles at the 27 cRM Y-STRs tested; these individuals and their fathers were excluded from all analyses. The mutation characteristics of each of the 27 cRM Y-STR marker are summarized in Table 1.

Table 1. Empirically established mutation rate estimates and mutation characteristics of 27 candidate RM Y-STR initially identified by our in silico approach, from genotyping 1,616 DNA confirmed father–son pairs.

Name	No. of father–son pairs genotyped	No. of mutations observed	Mutation rate ($\times 10^{-3}$)	95% Confidence interval ($\times 10^{-3}$)	Expansions (%)	Contractions (%)	*p*-value of direction	Unknown direction	1-Step (%)	2-Step (%)	3-Step (%)	Mutation rate category
DYF1001	1,616	84	52	[42, 64]	35 (42)	46 (55)	.266	3	79 (94)	4 (5)	0	RM
DYS724/ CDY	1,616	75	46	[37, 58]	34 (45)	41 (55)	.489	0	74 (99)	1 (1)	0 (0)	RM
DYF1000	1,616	58	36	[27, 46]	27 (47)	30 (52)	.791	1	n.a.[a]	n.a.[a]	n.a.[a]	RM
DYR88	1,616	47	29	[21, 39]	23 (49)	24 (51)	1.000	0	46 (98)	1 (2)	0 (0)	RM
DYS712	1,616	44	27	[20, 36]	26 (59)	18 (41)	.291	0	41 (91)	3 (7)	0 (0)	RM
DYS688/ DYS711	1,616	43	27	[19, 35]	25 (58)	18 (42)	.360	0	42 (98)	1 (2)	0 (0)	RM
DYS1012	1,616	31	19	[13, 27]	17 (55)	14 (45)	.720	0	29 (94)	2 (6)	0 (0)	RM
DYF1002	1,616	29	18	[12, 26]	15 (52)	14 (48)	1.000	0	29 (100)	0 (0)	0 (0)	RM
DYS1007	1,616	25	16	[10, 23]	12 (48)	12 (48)	1.000	1	25 (100)	0 (0)	0 (0)	RM
DYS1010	1,616	23	14	[9.0, 21]	10 (43)	13 (57)	.678	0	n.a.[a]	n.a.[a]	n.a.[a]	RM
DYS685/ DYS713	1,616	23	14	[9.0, 21]	12 (52)	11 (48)	1.000	0	21 (91)	1 (4)	1 (4)	RM
DYS1003	1,616	21	12	[7.1, 18]	4 (19)	16 (76)	**.012**	1	19 (90)	0 (0)	1 (5)	RM
DYS1013	1,616	16	9.9	[5.7, 16]	4 (25)	12 (75)	.077	0	15 (94)	1 (6)	0 (0)	FM
DYS1005	1,616	15	9.3	[5.2, 15]	8 (53)	7 (47)	1.000	0	15 (100)	0 (0)	0 (0)	FM
DYS1016	1,616	14	8.7	[4.7, 15]	9 (64)	5 (36)	.424	0	14 (100)	0 (0)	0 (0)	FM
DYS1017	1,616	13	8.0	[4.3, 14]	10 (77)	3 (23)	.092	0	13 (100)	0 (0)	0 (0)	FM
DYF1009	1,616	11	6.8	[3.8, 12]	6 (55)	5 (45)	1.000	0	10 (91)	0 (0)	1 (9)	FM
DYS1014	1,616	11	6.8	[3.4, 12]	6 (55)	5 (45)	1.000	0	11 (100)	0 (0)	0 (0)	FM
DYR33	1,616	11	6.8	[3.4, 12]	7 (64)	4 (36)	.549	0	11 (100)	0 (0)	0 (0)	FM
DYS714	1,616	10	6.2	[3.0, 11]	4 (40)	6 (60)	.754	0	9 (90)	1 (10)	0 (0)	FM
DYF1004	1,616	10	6.2	[3.0, 11]	4 (40)	5 (50)	1.000	1	9 (90)	0 (0)	0 (0)	FM
DYS1006	1,616	9	5.6	[2.5, 11]	7 (78)	2 (22)	.180	0	8 (89)	1 (11)	0 (0)	FM
DYS1015	1,616	8	5.0	[2.1, 9.7]	6 (75)	2 (25)	.289	0	8 (100)	0 (0)	0 (0)	MM
DYS563/ DYF408	1,616	6	3.7	[1.4, 8.1]	4 (67)	2 (33)	.688	0	6 (100)	0 (0)	0 (0)	MM

TABLE 1 (Continued)

Name	No. of father–son pairs genotyped	No. of mutations observed	Mutation rate ($\times 10^{-3}$)	95% Confidence interval ($\times 10^{-3}$)	Expansions (%)	Contractions (%)	*p*-value of direction	Unknown direction	1-Step (%)	2-Step (%)	3-Step (%)	Mutation rate category
DYF1011	1,616	5	3.1	[1.0, 7.2]	3 (60)	2 (40)	1.000	0	5 (83)	0 (0)	0 (0)	MM
DYS524/ DYF400	1,616	3	1.9	[0.4, 5.4]	0 (0)	3 (100)	.250	0	3 (100)	0 (0)	0 (0)	MM
DYS1008	1,616	2	1.2	[0.1, 4.5]	0 (0)	2 (100)	.500	0	2 (100)	0 (0)	0 (0)	MM
Overall	*1,616*	*647*	*15*	*[14, 16]*	*318 (49)*	*322 (50)*	*.906*	*7*	*544 (97)*	*16 (3)*	*3 (<1)*	

Note: Mutation rates and their associated confidence intervals are expressed as number of mutations per marker per meiosis. Novel Y-STRs and their newly proposed names (DYF/DYS10xx) are shown in italic.

Abbreviations: FM, fastly mutating; MM, moderately mutating; RM, rapidly mutating; STRs, short tandem repeats.

[a]The number of multistep mutations could not be assessed for this Y-STR marker as the sequence contained both trinucleotide repeat stretches and a hexanucleotide repeat stretch in DYF1000 and both tetranucleotide stretches and a dinucleotide stretch in DYF1010.

Following the mutation rate criteria described in method section 2.4, 12 (44%) out of the 27 cRM Y-STRs tested were classified as RM Y-STRs with mutation rate $>10^{-2}$, representing eight novel Y-STRs not previously described at all, and four Y-STRs previously described in population studies. The previously discovered Y-STRs were: DYS713 [44], later also described as DYS685 [45]; DYS711 [44], later also described as DYS688 [45]; DYS712 [44]; and CDY (included in commercial products of FamilyTreeDNA), later also described as DYS724 [46]. Three of those markers (DYS713, DYS711 and DYS712) had only population data and no mutation data previously reported [44, 47-49]. For one of the previously discovered Y-STR markers, DYS724, mutation data were previously inferred from population data [50] and later from deep-rooted pedigrees [23, 29], while mutation data from comprehensive father-son pair analysis as in the present study were not previously reported. Although not being described in scientific literature, another one of the newly classified RM Y-STRs is part of a test kit sold by a direct-to-consumer DNA testing company (i.e., FamilyTreeDNA) under the name DYR88.

Next to the identified 12 RM Y-STRs, the mutation rate data allowed classifying 10 of the 27 cRM Y-STR markers (37%) as FM Y-STRs with mutation rates between 5×10^{-3} and 1×10^{-2}, representing nine novel Y-STRs markers not previously described at all. One Y-STR markers was previously discovered [44], and population data were published: DYS714 [44, 48, 49]. One of the nine novel FM Y-STRs is also used by FamilyTreeDNA under the name: DYR33, but no marker information was found in scientific publications.

The remaining five cRM Y-STR markers (19%) were classified based on the mutation rate data as MM Y-STRs with mutation rates between 1×10^{-3} and 5×10^{-3}, representing three novel Y-STR markers not previously described at all, and two previously described Y-STR markers: DYS524 and DYS563 [51], which both lack population data and mutation rate data in the scientific literature. SM Y-STRs with mutation rates $<10^{-3}$ were not observed among the 27 cRM Y-STR markers tested, demonstrating the power of our *in silico* search strategy to find Y-STR markers with increased mutation rate. Notably, this is in contrast to our previous unbiased empirical screening study [11] that revealed 82 (44%) of 186 Y-STRs with mutation rates $<10^{-3}$.

Thus, overall, more than 80% of the cRM Y-STR markers highlighted via our *in silico* analysis designed to find Y-STRs with increased mutation rate were indeed empirically verified as Y-STRs with increased mutation rates, either RM Y-STRs or FM Y-STRs. This again contrasts markedly to the only 16% such markers i.e. 7% RM Y-STRs and

9% FM Y-STRs identified in our previous unbiased screening study, including 186 Y-STRs [11]. These results clearly demonstrate the advantage of applying our *in silico* approach, including the mutability prediction score, for identifying Y-STRs with increased mutation rates compared to the unbiased, massive screening approach applied previously [11]. In the present study, we applied our *in silico* approach only to the Y-chromosome reference sequence to identify Y-STRs with increased mutation rates. In the future, our *in silico* approach may also be applied to the autosomal reference sequence to identify autosomal STRs with increased mutation rates for suitable human genetic research and application purposes.

The set of newly identified 12 RM Y-STRs has a mean mutation rate of $2.6x10^{-2}$, which is higher compared to that of the set of previously identified 13 RM Y-STRs with $1.6x10^{-2}$ [17]. However, the most mutable of all currently known RM Y-STR markers remains one from the previously published set, namely DYF399S1, which has an estimated mutation rate of $6.9x10^{-2}$ [17]. In comparison, the most mutable novel RM Y-STR identified in the present study, DYF1001, has a slightly lower estimated mutation rate of $5.2x10^{-2}$. When combining the 12 novel with the 13 previous RM Y-STRs and ranking them according to their empirically derived mutation rate estimates, rank 2 – 6 go to newly identified RM Y-STRs, once again demonstrating the power of our combined *in silico* and empirical approach.

The newly identified RM Y-STR marker set contains slightly more multi-copy markers (five) compared to the previously published RM Y-STR set (four). It was not possible to separate the individual copies of such markers with our approach, therefore it remains unknown if the different copies contributed equally to the increased mutability of these markers. A total of ten out of the 27 cRM Y-STRs were multi-copy markers, of these ten only half were confirmed to be RM Y-STRs, therefore we can conclude that the increased mutability that stems from having multiple copies alone is not sufficient to explain the high mutability that can be found in some of these Y-STRs. Both RM Y-STR sets predominantly consist of tetranucleotide repeat loci; the previously published set contained only one trinucleotide repeat locus, while the newly identified set contains two trinucleotide loci (of which one also contains a hexanucleotide repeat). Note that homopolymers and dinucleotide repeats were not considered *a priori* in both the current and the previous study [11].

Besides the success of our *in silico* approach to identify novel RM Y-STRs, about half (56%) of the cRM Y-STRs highlighted *in silico* showed empirical mutation rates $<10^{-2}$ in the father-son pair testing, and thus were not empirically confirmed as RM Y-STR. This can be explained by various factors. One is the use of a strict mutation rate boundary of 10^{-2} for classifying RM Y-STRs, which means that a marker with a slightly lower mutation rate of e.g. $9.9x10^{-3}$ is not classified as RM Y-STR such as DYS1013 in the present study (Table 1). A second factor is the impact of stochastic effects that are inherently associated to STR mutability studies and that becomes more pronounced the lower the mutation rate is given sample size constrains, e.g., all ten FM Y-STRs found in this study have the RM Y-STR mutation rate boundary of 10^{-2} within their 95% confidence interval (Table 1). A third factor is the sole use of the human genome reference sequence to find cRM Y-STRs, which provides a hybrid Y-chromosome sequence of a small number of individuals only, which can never reflect Y-STR diversity in any human population. Thus, any population effect is ignored when using a single sequence in the candidate marker ascertainment as done here. For example, purely by chance, the reference genome may display a very long STR allele, while the majority of the individuals in a population carry shorter alleles. In such case, using father-son pair samples from such population for mutation rate estimation would thus reveal lower mutation rates than expected from the *in silico* analysis, given the known impact of Y-STR allele length on Y-STR mutation rates (see also below). Furthermore, mutability may be affected by other sequence structure based differences between the reference genome and the study population, that were not covered by our *in silico* approach.

An ideal STR mutability prediction model would use multiple reference sequences from individuals of multiple populations, or alternatively, use the median allele size obtained from genotyping of one or several populations. However, such an approach would require large (whole genome) sequencing datasets. Although such data sets are publically available, the vast majority of currently available sequencing data is produced by short read sequencing, which is not suitable for finding RM Y-STRs that contain relatively long and complex repetitive sequences [28]. In the future, accurate third generation sequencing technologies like Pacbio's Single Molecule, Real-Time (SMRT) Sequencing may help to overcome these limitations. The future analysis of high-quality, high-coverage, and long read whole genome sequences [52] may result in additional novel cRM Y-STR markers that should be tested in large numbers of father-son pairs to empirically establish their RM Y-STR status.

Male relative differentiation capacity

Using the full set of 27 cRM Y-STRs genotyped, a total of 518 (32%) of the 1,616 father-son pairs analyzed were differentiated by at least 1 Y-STR mutations. When only considering the 12 RM Y-STRs, a total of 424 (26%) father-son pairs were separated; of these, 352 (83%) pairs were differentiated by a single mutation, 66 (15%) by two mutations, 5 (1%) by three mutations, and a single pair (<1%) was separated by four mutations. It is not expected that the 32% father-son differentiation rate based on the total number of 27 cRM Y-STRs is biased, because these father-son pairs have not been used for marker discovery (which was solely based on the *in silico* approach). However, the 26% father-son differentiation rate for the 12 RM Y-STRs may reflect an overestimation, because the same father-son pair data were used for highlighting the 12 RM Y-STRs out of the 27 cRM Y-STRs. At this moment it is difficult to know how serious this overestimation is until empirical data from independent father-son pairs and other male relatives become available with future studies.

However, to get a first impression and to provide a theoretical expectation on how well these 12 novel RM Y-STRs differentiate paternally related men, we estimated male differentiation capacity by using the empirically derived mutation rate estimates from the current study for male relatives separated by one to ten meioses, and compared it to the estimates calculated in the same way for the 13 previously identified RM Y-STRs [11]. As evident from Figure 3, the set of 12 new RM Y-STRs provides somewhat higher male relative differentiation capacity within all groups of male relative when compared to the 13 previously known RM Y-STRs. Moreover, when combining all 25 RM Y-STRs, male relative differentiation capacity for all pairs of relatives were drastically increased with 44% of the father-son pairs (one meiosis), 69% of the brothers and grandfather-grandson pairs (two meioses), 83% of the uncle-nephews (three meioses) and 90% of the cousins (four meioses) being differentiated by at least one mutation, respectively. For paternal relatives separated by eight meioses and above, over 99% were differentiated with this set of 25 RM Y-STR markers. If future relative differentiation rates derived from empirical testing of independent samples can confirm these estimates, this will provide a significant boost in the practical application of RM Y-STRs for male relative differentiation, as highly relevant in forensic case work [3] and other fields such as genetic genealogy [4].

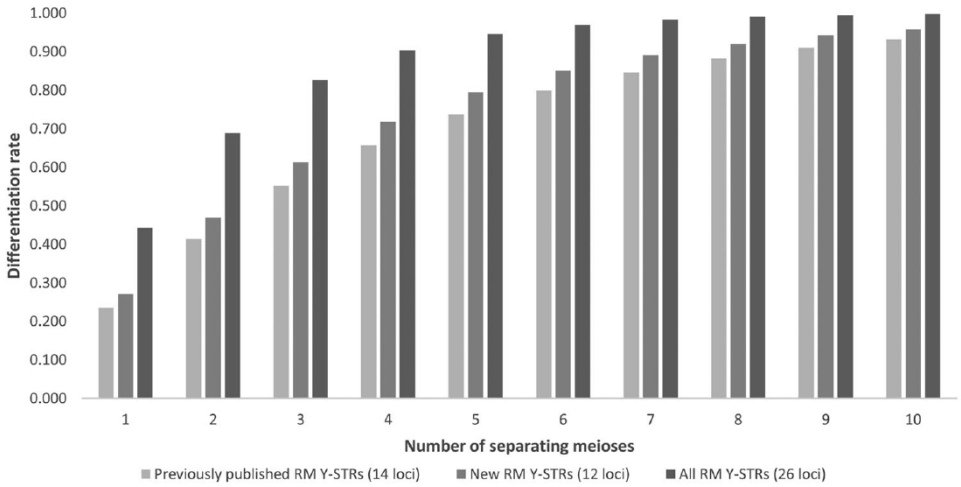


Figure 3: Male relative differentiation capacities calculated from the respective locus-specific mutation rate estimates for i) the 13 previously established RM Y-STRs [11], DYF403S1a and DYF403S1b were considered making a total of 14 loci. ii) the 12 novel RM Y-STRs identified in the present study, and iii) the combined set of 25 currently known RM Y-STRs, for male relative pairs separated by 1 to 10 meioses, respectively.

It is encouraging to note that for the 13 previously established RM Y-STRs, the mutation rate derived differentiation capacity estimates agreed well with the male relative differentiation rates empirically obtained from independent male relative data [17]. In particular, for pairs of men related by one to four meiosis, the differentiation capacity for the previous 13 RM Y-STRs were estimated to be 23%, 41%, 55% and 66%, respectively, while the empirically observed differentiation rates based on hundreds of relative pairs tested, were very similar at 24%, 44%, 55% and 61%, respectively [17]. Therefore it can be expected that provided enough male relative pairs being analyzed in future empirical studies, the empirically derived relative differentiation rates for the set of 12 novel RM Y-STRs and for the combined set of all 25 currently known RM Y-STRs shall be similar to the differentiation capacities presented here.

Internal and external factors influencing mutability

Impact of the length of the father's allele on Y-STR mutability

It is generally accepted that the length of an STR repeat, i.e. the number of repeats, is the most predominant driving factor of STR including Y-STR mutability [11, 28, 36-38, 40, 41].

Therefore it would be expected that fathers that possess long (Y-)STR alleles have an increased chance for a mutation to occur at these loci compared to fathers that possess short (Y-)STR alleles. Due to the relatively large number of 647 Y-STR mutations we observed at the 27 cRM Y-STRs among the >1,600 father-son pairs, we had the possibility to test this hypothesis for Y-STRs in particular. To this end, alleles observed in the fathers for each of the 27 cRM Y-STRs were classified as low, medium, or high length range alleles using the tertiles. The allelic mutation rates in each of the three categories were then calculated by dividing the total number of observed mutations by the total number of alleles and therefore represent the number of mutations per allele per meiosis. As shown in Figure 4, indeed the high range alleles with the longest repeats mutated more frequently than the low and the medium range alleles. There was a more than two-fold difference in allelic mutation frequency between the low and the high allele ranges. Pairwise comparison of proportions with conservative Bonferroni correction for multiple testing resulted in statistically significant p-values between all groups. The smallest difference was found between the low and medium allele ranges, with an adjusted p-value of 0.014, the adjusted p-value between the medium and high allele ranges was 1.1×10^{-9}, and between the low and high allele ranges the adjusted p-value was below 2×10^{-16}.

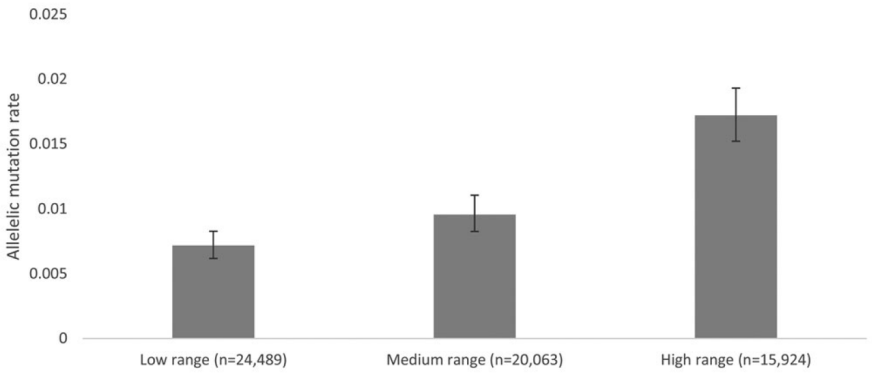


Figure 4: Y-STR allelic mutation rates (the number of mutations per allele per meiosis) of the genotyped 1,616 fathers according to the i) low, ii) medium and iii) high range allele groups (tertiles) as defined by the father's allelic fragment length based on the 27 cRM Y-STRs highlighted by our in silico approach.

It has also been previously suggested that some Y-STR markers may exhibit mutation rate differences between populations explained by different underlying Y-SNP haplogroups [29]. Theoretically, this could be caused, for instance under strong population

bottleneck scenarios involving a limited number of male founders, followed by (Y-chromosome) genetic isolation, when the male founders carry a predominant Y haplogroup associated with very short or very long Y-STR alleles instead of the more complete allele range the Y-STR would allow. In our study, Y haplogroup information was not available; but even if it were, it would be unlikely that this played a role in our study, given the German and Polish European descent of the father-son pairs used and their known Y haplogroup diversity [53]. However, it is encouraging that for most of the previously established set of 13 RM Y-STRs, the elevated mutation rates could be demonstrated in father-son pairs from different populations [13, 17, 18, 21, 54]. This suggests that the population and thus Y haplogroup background has a limited impact on the increased mutation rates of RM Y-STRs in most populations.

The directionality of mutations

Of the total of 647 observed mutations, the repeat expansion and contractions were nearly equally distributed with 318 expansions (49%) and 322 contractions (50%). To test if the direction of Y-STR repeat mutations was influenced by the allele length, we used the tertile based allele range grouping as described before. As seen in Figure 5, there appears to be a pattern where shorter alleles tend to expand more and the longer alleles contract more. Exact binomial testing showed a statistically significant difference in expansions and contractions in the low allele range, with more expansions than contraction (p-value 0.012), and a low, yet non-significant difference in the high allele range, with more contractions than expansions (p-value 0.061). In the medium allele range, however, the expansions and contractions appeared to be more balanced, as is also reflected in a non-significant p-value of 0.718. These results are in agreement with our previous study that found a similar effect of allele length on the direction of mutations across 186 Y-STRs [11]. The results are also in line with a study analyzing 236 mutations across 122 autosomal STRs, which demonstrated an exponential increase in the number of contractions with increasing allele size and predominantly expansion mutations in the lower allele size ranges [55].

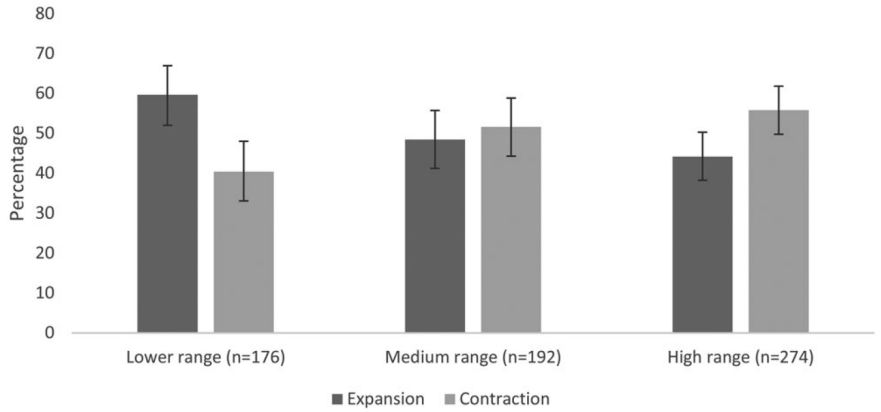


Figure 5: Y-STR repeat mutation expansion and contraction proportions according to the i) low, ii) medium and iii) high range allele groups as defined by the father's allelic fragment length. The groups were defined as the tertiles, based on 27 cRM Y-STRs highlighted by our in silico approach. The bars represent the binomial 95% confidence intervals.

Impact of the father's age on Y-STR mutability

Several previous studies showed that the father's age at time of siring his son affects STR including Y-STR mutability with a positive correlation; the older the father, the more mutations [11, 29, 56-58]. However, other studies reported no such, or only a small effect [59, 60], which may be explained by limited sample size effect or intrinsic differences (e.g. complexity, or sequence motifs) between the studied STRs. Taking advantage of the relatively large number of mutations we observed, we tested for the effect of father's age on the Y-STR mutability in our 27 cRM Y-STR markers.

To this end, all fathers of which the age at the time of conception was available (N=1,500) were divided in four groups defined by fathers' age at time of siring their sons according to the quartiles. We tested for outliers in the different age groups (individuals with age that fell outside of the range Q1 – 1.5 * IQR to Q3 + 1.5 *IQR), only two individuals (out of the 366) in the oldest age group could be considered outliers. As shown in Figure 6, indeed father's age had an impact on the number of observed mutations in our study. In the oldest age group there was a more than a twofold increase in the mean number of Y-STR mutations observed compared to the youngest age group. A pairwise comparisons using Wilcoxon rank sum test and applying Bonferroni p-value adjustment showed significant differences between the group with the largest number of Y-STR mutations: group 4 (oldest fathers) and all other age groups (p-values of 1.8×10^{-11}, 1.2×10^{-5}, and 0.0018 compared to group 1, 2, and 3, respectively). Additionally, the second oldest

age group 3 showed significantly more Y-STR mutations than the youngest age group 1 (p-value of 0.013), although the difference was much smaller than seen between group 4 and all other age groups. These results are in line with earlier observations of us and others that increased father's age increases (Y-)STR mutability [11, 29, 41, 56, 58]. Moreover, this finding highlights that when using father-son pairs to study (Y-)STR mutability, the age distribution of the fathers at the time of siring is a factor to consider when interpreting the mutation outcomes. Notably, although the average age that men become fathers has generally increased over the past decades for various reasons [61], there also are strong differences between populations based on various reasons including cultural and economic factors [62] that shall be considered for the data interpretation in future studies.

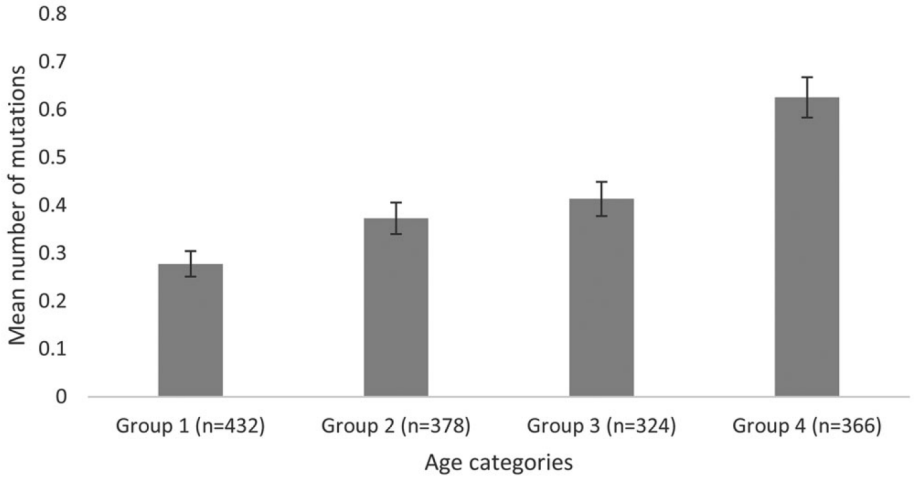


Figure 6: Mean number of observed Y-STR mutations according to four categories defined by the father's age at time of conception of his son. The age groups were defined as the quartiles. Group 1: 15 – 23 years old, Group 2: 24 – 29 years old, Group 3: 30 – 36 years old, and Group 4: 37 - 66 years old, based on 27 cRM Y-STRs highlighted by our in silico approach.

Impact of the repeat sequence motif on mutability

Based on previously published studies, it remains unclear if the DNA sequence of the repeat motif has a direct impact on the (Y-)STR mutability. Some studies described such effect [37, 38], while others did not see such [11]. Often it is difficult to study this effect, because STRs with different repeat motifs are typically not available in similarly large numbers, which may have to do with uneven distributions in the human genome and/or marker ascertainment due to study design. Our *in silico* approach did not consider repeat motifs in the marker ascertainment. However, in case the repeat motif positively impacts

on mutability, our *in silico* approach could reflect this, and thus would be biased, since we successfully (see above) enriched for markers with increased mutation rates. Testing for the effect of repeat motif sequence on Y-STR mutability using the 12 novel RM Y-STRs together with the 13 previously established RM Y-STRs, we observed a rather striking pattern when comparing them with 173 Y-STRs characterized by lower mutation rates (i.e. $<10^{-2}$). For this analysis, we considered repeat motif families, e.g., AAAT, AATA, ATAA, TAAA, TTTA, TTAT, TATT, and ATTT were all called as AAAT repeats family. For the 25 RM Y-STR markers we found that among the total of 34 tetranucleotide repeats (the different copies from multi-copy markers were considered as separate repeats here), 33 (97%) contained a repeat stretch belonging to the AAAG sequence motif family, and 12 (35%) contained a repeat stretch belonging to the AAGG sequence motif family (Table 2, Table S2). There was only one (3%) of the 34 tetranucleotide repeat RM Y-STR markers that did not contain either of those two motifs (DYS712), but instead consisted of a long AGAT and a short ACAG repetitive stretch. Similarly when focusing on the six trinucleotide repeats (derived from three RM Y-STR markers) among the 25 RM Y-STRs, all markers contained a repeat stretch belonging to the AAG sequence motif family and additionally half also contained an AGG sequence motif.

Table 2: Differences in observed STR sequence motifs between RM Y-STRs and non-RM Y-STRs. Significant p-values (Fisher's exact test) are shown in bold.

Motif	RM Y-STRs †	Non-RM Y-STRs ‡	p-value
[AAAG]	33 in 34	19 in 117	**<0.0001**
[AAGG]	12 in 34	22 in 117	0.0606
[AGAT]	1 in 34	37 in 117	**0.0003**
[AAAT]	1 in 34	37 in 117	**0.0003**
[AAG]	6 in 6	8 in 60	**<0.0001**
[AGG]	3 in 6	3 in 60	**0.0078**
[AAT]	0 in 6	34 in 60	**0.0100**
[AAC]	0 in 6	15 in 60	0.3234

† These represent a combinations of the 13 previously published RM Y-STRs (Ballantyne et al., 2010), and the 12 novel RM Y-STRs described in the present study.

‡ These represent non-RM Y-STRs (mutation rate < 10-2 mutations per marker per meiosis) from a previous study (Ballantyne et al., 2010).

In contrast, however, when assessing the motifs sequence families found in the 173 non-RM Y-STR markers from the Ballantyne et al. (2010) study, among the 117

tetranucleotide repeats the AAAG and AAGG motif families were only found in 16% and 19%, of the repeats respectively (Table S2), which is considerably lower than we found for the RM Y-STRs (p-values <0.0001 and 0.0606, respectively, Table 2). The most frequently observed tetranucleotide motif sequences in these non-RM Y-STR loci belonged to the AAAT and AGAT repeat sequence families, both found in 32% of these non-RM STRs (Table 2, Table S2). In contrast, both the AAAT and the AGAT sequence motif families were found only once among the 34 tetranucleotide RM Y-STR loci (p-value 0.0003 in both cases). Similarly, among the 60 trinucleotide non-RM Y-STR loci from Ballantyne et al. [11], the AAG and AGG sequence motif families were found only in 13% and 5%, respectively (p-value of <0.0001 and 0.0078, respectively, Table 2, Table S2), while their most frequently observed motifs were AAT and AAC at 57% and 25%, respectively (Table 2, Table S2), which were completely absent in the six trinucleotide RM Y-STR loci (p-value 0.0100 and 0.3234, respectively).

Although the total number of RM Y-STRs available for this analysis is relatively small, and consequently the number of tetranucleotide and trinucleotide RM Y-STRs, our findings suggest that there are statistically significant differences in sequence motif depending on the mutation rate of the Y-STRs i.e., between RM Y-STRs and non-RM Y-STRs (Table 2). In turn, these results would allow concluding an impact of repeat sequence motif on (Y-)STR mutability in line with some previous studies [37, 38]. One explanation may be the formation of secondary structures, in particular triplex DNA, which can by formed by homopurine repeat motifs (e.g. AAG, AGG, AAAG, AAGG) [38, 63, 64]. Whether, this would affect the mutability directly, or rather impacts on the direction of mutations [65], leading to longer repeat stretches and thus a higher mutability, remains to be understood in future more dedicated studies. The STR structure sequences of all RM Y-STRs and non-RM Y-STRs used in this analysis can be found in the supplementary materials (Table S2).

Conclusions

We developed and provide a novel *in silico* method to find STRs with increased mutation rates from searching sequencing data, which in the future can be applied for all types of research questions for which highly mutable STRs are required. The application of the *in silico* method to the human reference sequence by focusing on the Y-chromosome allowed us to highlight 27 candidate RM Y-STR markers, for which subsequent empirical testing in 1,616 DNA-confirmed father-son pairs identified 12 novel RM Y-STRs (mutation rate $>10^{-2}$) and 11 novel FM Y-STRs (mutation rate 5×10^{-3}-10^{-2}). We showed that the 12

novel RM Y-STRs outperform the 13 previously identified RM Y-STRs in male relative differentiation capacity, and that the combined set of 25 RM Y-STRs provides strongly increased male relative differentiation capacity compared to both separate sets, which will need to be confirmed in future studies to establish empirical male relative differentiation rates. The large number of 647 Y-STR mutations we observed allowed us to establish internal and external factors such as the length of the allele and the age of the father at the time of conception to impact on Y-STR mutability. Overall, we expect that the 12 novel RM Y-STRs identified in the present study, in combination with the 13 RM Y-STRs we identified previously, will allow significantly improving the differentiation ability of paternally related men, close as well as distant ones, in future human genetic applications such as in forensic casework and genealogical studies.

Data Availability Statement

The data that support the findings of this study are available for scientific research purpose from the corresponding author upon reasonable request.

Acknowledgements

The authors would like to thank Diego Montiel González and Dion Zandstra for technical assistance with computational and coding matters during the data analysis, and Benjamin Planterose Jiménez for statistical advice. The work of AR, NK and MK was supported by Erasmus MC University Medical Center Rotterdam.

Conflict of interest

AR and MK are inventors of a filed patent application “Novel Y-chromosomal short tandem repeat markers for typing male individuals” (EP20158807).

References

1. Lygo, J.E., et al., *The validation of short tandem repeat (STR) loci for use in forensic casework.* International Journal of Legal Medicine, 1994. **107**(2): p. 77-89.
2. Fregeau, C.J. and R.M. Fourney, *DNA typing with fluorescently tagged short tandem repeats: a sensitive and accurate approach to human identification.* Biotechniques, 1993. **15**(1): p. 100-119.
3. Kayser, M., *Forensic use of Y-chromosome DNA: a general overview.* Human Genetics, 2017. **136**(5): p. 621-635.
4. Calafell, F. and M.H.D. Larmuseau, *The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research.* Human Genetics, 2017. **136**(5): p. 559-573.
5. Jobling, M.A. and C. Tyler-Smith, *Human Y-chromosome variation in the genome-sequencing era.* Nature Reviews Genetics, 2017. **18**(8): p. 485.
6. Roewer, L., *Y chromosome STR typing in crime casework.* Forensic Science, Medicine, and Pathology, 2009. **5**(2): p. 77-84.
7. Vuichard, S., et al., *Differential DNA extraction of challenging simulated sexual-assault samples: a Swiss collaborative study.* Investigative Genetics, 2011. **2**(1): p. 11.
8. Gill, P., et al., *Genotyping and interpretation of STR-DNA: low-template, mixtures and database matches—twenty years of research and development.* Forensic Science International: Genetics, 2015. **18**: p. 100-117.
9. Phillips, C., *The Golden State Killer investigation and the nascent field of forensic genealogy.* Forensic Science International: Genetics, 2018. **36**: p. 186-188.
10. Claerhout, S., et al., *Ysurnames? The patrilineal Y-chromosome and surname correlation for DNA kinship research.* Forensic Science International: Genetics, 2020. **44**: p. 102204.
11. Ballantyne, K.N., et al., *Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications.* The American Journal of Human Genetics, 2010. **87**(3): p. 341-353.
12. Ballantyne, K.N., et al., *A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages.* Forensic Science International: Genetics, 2012. **6**(2): p. 208-218.
13. Ballantyne, K.N., et al., *Toward Male Individualization with Rapidly Mutating Y-Chromosomal Short Tandem Repeats.* Human Mutation, 2014. **35**(8): p. 1021-1032.
14. Alghafri, R., W. Goodwin, and S. Hadi, *Rapidly mutating Y-STRs multiplex genotyping panel to investigate UAE population.* Forensic Science International: Genetics Supplement Series, 2013. **4**(1): p. e200-e201.
15. Robino, C., et al., *Development of an Italian RM Y-STR haplotype database: results of the 2013 GEFI collaborative exercise.* Forensic Science International: Genetics, 2015. **15**: p. 56-63.
16. Westen, A.A., et al., *Analysis of 36 Y-STR marker units including a concordance study among 2085 Dutch males.* Forensic Science International: Genetics, 2015. **14**: p. 174-181.
17. Adnan, A., et al., *Improving empirical evidence on differentiating closely related men with RM Y-STRs: a comprehensive pedigree study from Pakistan.* Forensic Science International: Genetics, 2016. **25**: p. 45-51.
18. Boattini, A., et al., *Mutation rates and discriminating power for 13 rapidly-mutating Y-STRs between related and unrelated individuals.* PLOS One, 2016. **11**(11): p. e0165678.
19. Niederstätter, H., et al., *Differences in urbanization degree and consequences on the diversity of conventional vs. rapidly mutating Y-STRs in five municipalities from a small*

region of the Tyrolean Alps in Austria. Forensic Science International: Genetics, 2016. **24**: p. 180-193.
20. Turrina, S., et al., *Are rapidly mutating Y-short tandem repeats useful to resolve a lineage? Expanding mutability data on distant male relationships.* Transfusion, 2016. **56**(2): p. 533-538.
21. Lang, M., et al., *Comprehensive mutation analysis of 53 Y-STR markers in father-son pairs.* Forensic Science International: Genetics Supplement Series, 2017. **6**: p. e57-e58.
22. Zgonjanin, D., et al., *Rapidly mutating Y-STRs population data in the population of Serbia and haplotype probability assessment for forensic purposes.* Forensic Science International: Genetics Supplement Series, 2017. **6**: p. e383-e384.
23. Boattini, A., et al., *Estimating Y-Str Mutation Rates and Tmrca Through Deep-Rooting Italian Pedigrees.* Scientific Reports, 2019. **9**(1): p. 9032.
24. Salvador, J.M., et al., *Filipino DNA Variation at 36 Y-chromosomal Short Tandem Repeat (STR) Marker Units.* Philippine Journal of Science, 2019. **148**: p. 43-52.
25. Larmuseau, M.H.D., et al., *A Historical-Genetic Reconstruction of Human Extra-Pair Paternity.* Current Biology, 2019. **29**(23): p. 4102-4107. e7.
26. Roewer, L., *Y-chromosome short tandem repeats in forensics—Sexing, profiling, and matching male DNA.* Wiley Interdisciplinary Reviews: Forensic Science, 2019. **1**(4): p. e1336.
27. Goedbloed, M., et al., *Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFlSTR® Yfiler® PCR amplification kit.* International Journal of Legal Medicine, 2009. **123**(6): p. 471.
28. Willems, T., et al., *Population-scale sequencing data enable precise estimates of Y-STR mutation rates.* The American Journal of Human Genetics, 2016. **98**(5): p. 919-933.
29. Claerhout, S., et al., *Determining Y-STR mutation rates in deep-routing genealogies: identification of haplogroup differences.* Forensic Science International: Genetics, 2018.
30. Larmuseau, M.H.D., et al., *Low historical rates of cuckoldry in a Western European human population traced by Y-chromosome and genealogical data.* Proceedings of the Royal Society B: Biological Sciences, 2013. **280**(1772): p. 20132400.
31. Claerhout, S., et al., *A game of hide and seq: Identification of parallel Y-STR evolution in deep-rooting pedigrees.* European Journal of Human Genetics, 2019. **27**(4): p. 637.
32. Benson, G., *Tandem repeats finder: a program to analyze DNA sequences.* Nucleic Acids Research, 1999. **27**(2): p. 573-580.
33. Hauge, X.Y. and M. Litt, *A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR.* Human Molecular Genetics, 1993. **2**(4): p. 411-415.
34. Mensah, M.A., et al., *Pseudoautosomal region 1 length polymorphism in the human population.* PLOS Genetics, 2014. **10**(11): p. e1004578.
35. Poriswanish, N., et al., *Recombination hotspots in an extended human pseudoautosomal domain predicted from double-strand break maps and characterized by sperm-based crossover analysis.* PLOS Genetics, 2018. **14**(10): p. e1007680.
36. Ellegren, H., *Microsatellites: simple sequences with complex evolution.* Nature Reviews Genetics, 2004. **5**(6): p. 435.
37. Kelkar, Y.D., et al., *The genome-wide determinants of human and chimpanzee microsatellite evolution.* Genome Research, 2008. **18**(1): p. 30-38.
38. Eckert, K.A. and S.E. Hile, *Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome.* Molecular Carcinogenesis, 2009. **48**(4): p. 379-388.
39. Kayser, M., et al., *A comprehensive survey of human Y-chromosomal microsatellites.* The American Journal of Human Genetics, 2004. **74**(6): p. 1183-1197.

40. Kayser, M., 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.* The American Journal of Human Genetics, 2000. **66**(5): p. 1580-1588.
41. Brinkmann, B., et al., *Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat.* The American Journal of Human Genetics, 1998. **62**(6): p. 1408-1415.
42. Tusnady, G.E., et al., *BiSearch: primer-design and search tool for PCR on bisulfite-treated genomes.* Nucleic Acids Research, 2005. **33**(1): p. e9-e9.
43. Vallone, P.M. and J.M. Butler, *AutoDimer: a screening tool for primer-dimer and hairpin structures.* Biotechniques, 2004. **37**(2): p. 226-231.
44. Leat, N., et al., *Properties of novel and widely studied Y-STR loci in three South African populations.* Forensic Science International, 2007. **168**(2-3): p. 154-161.
45. Maybruck, J.L., et al., *A comparative analysis of two different sets of Y-chromosome short tandem repeats (Y-STRs) on a common population panel.* Forensic Science International: Genetics, 2009. **4**(1): p. 11-20.
46. Jacobs, M., et al., *Development and evaluation of multiplex Y-STR assays for application in molecular genealogy.* Forensic Science International: Genetics Supplement Series, 2009. **2**(1): p. 57-59.
47. Hanson, E.K. and J. Ballantyne, *Population data for 48 'Non-Core'Y chromosome STR loci.* Legal Medicine, 2007. **9**(4): p. 221-231.
48. Zhang, G.Q., et al., *Structure and polymorphism of 16 novel Y-STRs in Chinese Han Population.* Genetics and Molecular Research, 2012. **11**(4): p. 4487-4500.
49. Liu, J., et al., *Development of a new 17 Y-STRs system using fluorescent-labelled universal primers and its application in Shanxi population in China.* Forensic Science International: Genetics Supplement Series, 2019. **7**(1): p. 95-97.
50. Chandler, J.F., *Estimating per-locus mutation rates.* Journal of Genetic Genealogy, 2006. **2**: p. 27-33.
51. Hanson, E.K. and J. Ballantyne, *Comprehensive annotated STR physical map of the human Y chromosome: Forensic implications.* Legal Medicine, 2006. **8**(2): p. 110-120.
52. Vollger, M.R., et al., *Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads.* BioRxiv, 2019: p. 635037.
53. Kayser, M., et al., *Significant genetic differentiation between Poland and Germany follows present-day political borders, as revealed by Y-chromosome analysis.* Human Genetics, 2005. **117**(5): p. 428-443.
54. Zgonjanin, D., et al., *Mutation rate at 13 rapidly mutating Y-STR loci in the population of Serbia.* Forensic Science International: Genetics Supplement Series, 2017. **6**: p. e377-e379.
55. Xu, X., et al., *The direction of microsatellite mutations is dependent upon allele length.* Nature Genetics, 2000. **24**(4): p. 396.
56. Sun, J.X., et al., *A direct characterization of human mutation based on microsatellites.* Nature Genetics, 2012. **44**(10): p. 1161.
57. Kong, A., et al., *Rate of de novo mutations and the importance of father's age to disease risk.* Nature, 2012. **488**(7412): p. 471.
58. Gusmao, L., et al., *Mutation rates at Y chromosome specific microsatellites.* Human Mutation, 2005. **26**(6): p. 520-528.
59. Forster, P., et al., *Elevated germline mutation rate in teenage fathers.* Proceedings of the Royal Society B: Biological Sciences, 2015. **282**(1803): p. 20142898.
60. Dupuy, B.M., et al., *Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci.* Human Mutation, 2004. **23**(2): p. 117-124.
61. Khandwala, Y.S., et al., *The age of fathers in the USA is rising: an analysis of 168 867 480 births from 1972 to 2015.* Human Reproduction, 2017. **32**(10): p. 2110-2116.

62. Young Jr, A.A., *Comment: Reactions from the perspective of culture and low-income fatherhood.* The ANNALS of the American Academy of Political and Social Science, 2011. **635**(1): p. 117-122.
63. Slebos, R.J.C., et al., *Mutations in tetranucleotide repeats following DNA damage depend on repeat sequence and carcinogenic agent.* Cancer Research, 2002. **62**(21): p. 6052-6060.
64. Zhao, J., et al., *Non-B DNA structure-induced genetic instability and evolution.* Cellular and Molecular Life Sciences, 2010. **67**(1): p. 43-62.
65. Shah, S.N., S.E. Hile, and K.A. Eckert, *Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes.* Cancer Research, 2010. **70**(2): p. 431-435.

Supplementary data

Table S1: Overview of the candidate RM Y-STRs and some of their characteristics.

Published name*	Proposed name*	Mutability predication score	Genomic amplicon range(s) (GRCh38)	Forward primer	Reverse primer	C.E. size range	Motif size	Single / multi copy	Complexity	Multiplex number
n.a.	DYF1000		chrY:17854578-17854819; chrY:18051088-18051432; chrY:24024183-24024480; 123 chrY:25645653-25645953 chrY:23054609-23059044; chrY:24688349-24688790; 107 chrY:24981317-24981750 chrY:24005865-24006148;	AGGGAGCTTCAGTGTGCTTC	TGGCTTGACTCAGAAACTTGAC	250-360	3	3 Multi-copy	Complex	6
n.a.	DYF1001			GCCTGGGCAACAAAAGTG	GCACAGCAGACCTGTTCTACAA	380-444	4	4 Multi-copy	Complex	2
DYS724/CDY[a]	n.a.		69 chrY:25663987-25664275	GGACTTAAGGAATTTCAACTCAT	GGCTCATTGTAGTCCACACTTGA	260-325	4	4 Multi-copy	Complex	1
DYS688[b]/DYS711[b]	n.a.		68 chrY:8465358-8465755	TATTCTAGACTCTGCTCAGAGAGGTG	GCTGTCATTGTATCTCTTCCACTCC	360-420	3	3 Single-copy	Complex	6
n.a.	DYF1002		48 chrY:16342720-16343155 chrY:24331827-24332233;	GCAAGACCATGTCTTGAAAG	GCAAGTCCTCATTATTTGC	382-430	4	4 Multi-copy	Complex	2
DYR88[c]	n.a.		35 chrY:25337916-25338322	CACCTGTAATCCTAGCTACTCAGA	GCTTTCTCATCATAGAGTGCTCATG	370-446	4	4 Multi-copy	Simpel	1
DYS720[d] (partial overlap)	DYS1003		34 chrY:13508603-13508009 chrY:18466997-18467397;	GTGAGACTCCATCTCAAAAGAAG	ATCTGGTGTGTGAAGGACAGC	348-406	4	4 Single-copy	Complex	1
n.a.	DYF1004		30 chrY:18852999-18853399 chrY:17642914-17643292;	GCAAAGCAGAATATTTATGTTTCAGT	AAAAATTTGCGAGGTGTGG	382-408	4	4 Multi-copy	Complex	4
n.a.	DYF1009		26 chrY:18262718-18263092	TCCCAGAGTCTCAGAGTGCTG	GCCATAGAAGTCACCAGCTAGG	354-384	4	4 Multi-copy	Complex	3
n.a.	DYS1005		25 chrY:7849377-7849841	CTATAGTTCCTAGAGCAAAGAGTGAGC	AGGTTGTGGTAATCTGAGATTGC	450-510	4	4 Single-copy	Complex	4
DYS685[b]/DYS713[b]	n.a.		25 chrY:7963956-7964261	GGGCTTTATAAGTTATCTGAGGC	GGTGACAGACTCCATCTTAA	294-330	4	4 Single-copy	Complex	2
DYS710[d] (partial overlap)	DYS1010		24 chrY:17324747-17325139 chrY:17787597-17787931;	AGAATCTATGATCACAGCACTTTTCT	CTTTCTACCTCATTTTCTTTCTGTCC	356-398	4	4 Single-copy	Complex	4
DYS524[e]/DYF400	n.a.		23 chrY:18118078-18118412	GCCAAGATCACAACCATTG	TGTATGAGCAATGCTAAGTTAAAACAA	312-338	4	4 Multi-copy	Complex	4
n.a.	DYS1007		23 chrY:17299905-17300373	GCACATGTCTATTAGTCCCATCTACC	GCCACTGAGTAATTCTTGAGAATTT	446-482	4	4 Single-copy	Complex	3
n.a.	DYS1012		23 chrY:5685298-5685581 chrY:22690460-22690791; chrY:23505396-23505720;	GCAAAGATCTCCATCTCCAAAAAG	CAAGCTTGGGTCCATTATGA	238-300	4	4 Single-copy	Complex	5
n.a.	DYF1011		23 chrY:26164819-26165213	TCCAATCAGGAACATTTTAAAGC	TAAGTTATCTTTAGCTTTTTGCTGGC	302-326	4	4 Multi-copy	Simpel	3
n.a.	DYS1008		21 chrY:7443263-7443654	ACATGATAGGAAATATGATTCTGGG	AAATTAGTTGGAAGTGTGTGACCCT	356-396	4	4 Single-copy	Complex	5
n.a.	DYS1016		21 chrY:12092576-12093208	GAGTGAAACCATCTCTAAGAC	GCTTGCAAACACACCATGA	588-616	4	4 Single-copy	Complex	2
n.a.	DYS1015		19 chrY:15682003-15682475	GCATATCATCTATCACCTATATTAT	ATAAATAGAAGATAGGTAGATAAC	440-498	4	4 Single-copy	Complex	2
n.a.	DYS1006		18 chrY:13208543-13208915	CAAGACCAGCATATCAGTTGAGG	GGCTCACTGCAAACCTCCTAC	346-372	4	4 Single-copy	Complex	6
DYS505[f] (partial overlap)	DYS1014		18 chrY:3772704-3773174	ACCTTCTGGCGAAGTAACCC	TCTCAGCACTGCACTCTAGCT	454-514	4	4 Single-copy	Complex	4
n.a.	DYS1013		17 chrY:14697144-14697687	TTATCTTGCTGACCATTAGAGAAGAT	CCTCTACCTACCTAGCTATGTA	530-560	4	4 Single-copy	Complex	5
DYS712[b]	n.a.		15 chrY:13446462-13446882	GTGGCTCACTCCTATAACC	AACTTTGTAGTAAGTTTTGAGACTGG	392-456	4	4 Single-copy	Complex	1
DYR33[b]	n.a.		13 chrY:17388667-17388955	ACTGAACGCTAGCCTGGGT	ACCCATATAGCTATCCATCTGTCTAAT	250-300	4	4 Single-copy	Complex	5
n.a.	DYS1017		13 chrY:5541615-5541844 chrY:17911925-17912161;	ATAGGACAGGTGGTTGAATG	CTAGATTTTCAGATAGTACAACC	200-234	4	4 Single-copy	Simpel	2
DYS563[f]/DYF408	n.a.		11 chrY:17993821-17994081	GCCTGAGTAACACTGCAAGATTCT	AGAGACAATTTCTATGCTCCAGC	228-266	4	4 Multi-copy	Simpel	3
DYS714[d]	n.a.		7 chrY:19985806-19986053	ACCGAAGCTGTTCCTATTAGGC	CTGTCAAAGATTGATAATGGGTATGGTAC	214-286	5	5 Single-copy	Simpel	6

Table S1 (continued)

HGVS based repeat stucture description
NC_000024.10:g.17854669_17854886AAG[6]GAGAAG[5]GAG[1]AAG[3]GAG[4]AAG[5]GAGG[1]AAG[6]GAGAG[1]AAG[2]G[1]AAG[4]ATGAAGA[1]AAG[26];
NC_000024.10:g.18051120_18051340TTC[26]T[1]TTC[2]ATC[1]TTC[4]C[1]TTC[2]TCTCC[1]TTC[6]CTCC[1]TTC[6]TCC[4]TTC[3]TCCTTC[5]TCC[1]TTC[5];
NC_000024.10:g.24024277_24024447AAG[9]CAG[1]AAG[22]AGGAAGAGG[1]AAG[22];
NC_000024.10:g.25645685_25645858TTC[22]CTCTTCCTC[1]TTC[24]TGC[1]TTC[8]
NC_000024.10:g.23054646_23055003AAGG[9]AGAG[1]AAAG[3]AC[1]AAAG[15]AGAAAGAGAG[1]AAAG[4]AGAAAG[1]AG[4]N[24]AAGG[4]AAG[1]AAAG[13]AAAAGA[1]AAAG[4]AAGG[3]AAAGAG[1]AAAG[3]G[1]AAGG[14];
NC_000024.10:g.24688386_24688749AAGG[12]AGAG[1]AAAG[3]AC[1]AAAG[16]AGAG[1]AAAG[4]AGAAAG[1]AG[4]N[24]AAGG[6]AAG[1]AAAG[11]AAATAAAAGA[1]AAAG[4]AAGG[3]AAAGAG[1]AAAG[3]GAAGA[1]AAAG[11];
NC_000024.10:g.24981356_24981711TTCC[10]TTTCTTCC[1]TTTC[3]TCT[1]TTCC[3]TTC[1]TTTC[3]TTTTCTTTTA[1]TTTC[14]TTCC[6]TTCTT[1]TTTC[2]ATTTTCTCTCTT[1]TC[5]TTTCTC[1]TTTC[4]TCTC[1]TTTC[14]TTTGTC[1]TTTC[3]TCTCC[1]TTCC[9]
NC_000024.10:g.24005905_24006099CTTT[2]TTT[1]CTTT[3]CTCT[1]CTTT[2]CT[5]C[1]CCTT[16]CCTTTCCTT[1]CTTT[19];
NC_000024.10:g.25664038_25664236AAAG[21]AAGGA[1]AAGG[16]GA[5]G[1]AAAG[2]AGAG[1]AAAG[3]AAA[1]AAAG[2]
NC_000024.10:g.8465426_8465675CTT[5]T[1]CTT[30]CTCCTCCTCCTTCTC[1]CTT[3]CTCCTT[4]CTCCTA[1]CTT[25]CTCCTC[1]CTT[3]
NC_000024.10:g.16242254_16242529GA[7]GAAA[4]TA[1]GAAA[13]GA[1]GAAA[4]G[1]GAAA[13]CAGA[1]GAAA[4]GA[1]GAAA[10]AAGAAAAGAAAGAGGAA[1]GAAA[4]AAGA[1]GAAA[2]GG[1]AGGG[3];
NC_000024.10:g.16342793_16343084TCCC[3]TCC[1]TTTC[2]TCTT[1]TTTC[4]TTCCTCTTTCTTTTCTT[1]TTTC[13]TC[1]TTTC[4]TCTG[1]TTTC[2]TTC[1]TTTC[10]C[1]TTTC[5]CC[1]TTTC[13]TA[1]TTTC[4]TC[7]
NC_000024.10:g.24331961_24332037AAAG[19];
NC_000024.10:g.25338114_25338188CTTT[19]
NC_000024.10:g.15084645_15084969AAAG[4]GAACG[1]AAGG[11]AAG[1]AAAG[2]AG[1]AAAG[2]AAAAAAG[1]AGGG[3]AAGGAGGAAGGAAGGAAAG[1]AG[5]GGAGAGAGAACGGAAG[1]AAAG[18]GTAG[1]AAAG[14]AG[1]AAAG[4]AA[1]AAAG[3]AA[1]AAAG[2]
NC_000024.10:g.18467124_18467223TCCT[17]CCCT[4]CT[8];
NC_000024.10:g.18853175_18853273GA[8]GGGA[3]GGG[1]AGGA[17]
NC_000024.10:g.17643029_17643147TC[4]TTTC[3]CCTT[15]CATTTCTTCCTC[1]TCTT[14];
NC_000024.10:g.18262864_18262978AAGA[7]GAGGAAGAAAT[1]GAAG[14]GAAA[3]GA[4]
NC_000024.10:g.7849517_7849773CTTTC[11]CTTTCCT[1]TCCC[7]TTCTTTCT[1]CTTT[5]CCTT[13]CTTT[12]CTT[1]CCTT[4]CTTT[2]
NC_000024.10:g.7964032_7964196 CTTT[18]CT[1]CTTT[2]CTGTCTTTTTCTTTCT[1]CTTT[12]TT[1]CTTT[4]
NC_000024.10:g.17324793_17325010AAAG[17]AG[13]AAAG[11]AAAAA[1]AG[3]GGAGG[1]GA[3]GTATGAGGA[1]AGGG[3]AGTATGAGGA[1]AGGG[3]AGC[1]AAGG[3]
NC_000024.10:g.17787707_17787797AAGG[8]AAG[1]AAGG[14];
NC_000024.10:g.18118211_18118301TTCC[14]TTC[1]TTCC[8]
NC_000024.10:g.17300035_17300228AAAG[14]AG[1]AAAG[3]AG[1]AAAG[5]AG[1]AAAG[14]N[12]AG[5]N[12]AG[5]
NC_000024.10:g.56858316_56858541AAAG[15]GAAAGA[1]AAAG[2]A[1]AAAG[3]A[1]AAAG[5]A[1]AAAG[3]AAGGAAAG[4]AAGGAAGG[1]AAAG[16]
NC_000024.10:g.22690566_22690759AAGG[2]AAAGAAAGAAGGAGGG[1]AAGG[15]AGGAAGAAGA[1]AAGG[2]AGGAAGGAAGACAG[1]AAGG[3]N[35]AAGG[2]AAGAAAGGAAA[1]AAGG[3];
NC_000024.10:g.23505427_23505625TTCC[3]TTTTTCCTTTC[1]TTCC[3]TCCCTCCCTTTCTTCCTTC[1]TTCC[2]TC[3]CC[1]TTCC[2]TTCTGTC[1]TTCC[2]N[21]TTCC[13]CTCCTTCTTTCT[1]TTCC[2]TTCT[1]TTCC[2];
NC_000024.10:g.26165000_26165181AAGG[2]AAAGAAAGAAGGAGGG[1]AAGG[12]AGGAAGAAGA[1]AAGG[2]AGGAAGGAAGACAG[1]AAGG[3]N[35]AAGG[2]AAGAAAGGAAA[1]AAGG[3]
NC_000024.10:g.7443295_7443613TTCC[11]TTAT[1]TTCC[6]TTAT[1]TTCC[4]TCCCACCTTT[1]CCTC[3]CTTTCTTTCTCT[1]TTCC[9]TTAT[1]TTCC[4]TCCCACCTTT[1]CCTC[3]CTTTCTTTCTCT[1]TTCC[14]TCCC[3]A[1]CCTC[3]CTTCCTTT[1]CT[7]
NC_000024.10:g.12092653_12093169AAGG[4]AGTG[1]AGGG[5]AAAAAAGAAAGAAGG[1]GA[4]GGGAGAAAGAG[1]AAGG[2]AGAGTAGGAAA[1]AAGG[3]N[110]AGGG[3]N[29]AAGG[4]N[51]AAGG[6]N[21]AAAC[3]N[63]AAAG[13]AGAAAGAAAG[1]AAGG[2]A[1]AAGG[4]A[1]AAGG[4]AAAG[3]
NC_000024.10:g.15682183_15682426TTCC[4]TCC[1]TTCC[3]TTCTTTCCTTCCCTC[1]CT[5]C[1]TTTC[2]TCCCTCTG[1]TTTC[2]TC[1]TTTC[3]T[1]TTTC[12]TC[1]TTTC[2]TC[1]TTTC[4]TC[1]TTTC[4]GTTC[1]TTTC[9]TC[1]TTTC[3]
NC_000024.10:g.13208631_13208837AAG[9]AGG[1]AAG[2]AGG[1]AAG[4]GGGGAGAAGAAG[4]GGGGAG[1]AAG[5]AGGAAG[3]AGG[1]AAG[22]
NC_000024.10:g.3772790_3773129CCTT[12]N[56]TTTC[2]N[44]TTTCC[11]TCTCTTTCTCTT[1]TC[3]TTTCTCTTCC[1]TTCCC[3]TTTCCTCTC[1]TTTC[12]TT[1]TTTC[3]TT[1]TTTC[3]
NC_000024.10:g.14697183_14697634AGAT[2]AGGTAGGTAGGC[1]AGAT[11]N[81]AGAT[2]AGAA[1]AGAT[6]N[128]AGAT[10]N[59]AGAT[10]
NC_000024.10:g.13446552_13446639AGAT[17]AGAC[5]
NC_000024.10:g.17388716_17388923AAAG[15]N109AGAT[5]GAT[1]AGAT[4]
NC_000024.10:g.5541691_5541758CTTT[17]
NC_000024.10:g.17912088_17912119AGAT[8];
NC_000024.10:g.17993864_17993919ATCT[14]
NC_000024.10:g.19985848_19985982CTTTT[18]N[30]CTTTT[3]

* In cases where the new candidate RM Y-STR partially overlapped with an already known Y-STR, a new name was proposed to describe the marker in its entirety

a Maybruck, Julie L., et al. "A comparative analysis of two different sets of Y-chromosome short tandem repeats (Y-STRs) on a common population panel." Forensic Science international: genetics 4.1 (2009): 11-20.

b Leat, Neil, et al. "Properties of novel and widely studied Y-STR loci in three South African populations." Forensic science international 168.2-3 (2007): 154-161.

c FamilyTreeDNA

d Discovery unknown due to discontinuation of GDB, no scientific literature regarding the discovery could be found

e Hanson, Erin K., and Jack Ballantyne. "Comprehensive annotated STR physical map of the human Y chromosome: Forensic implications." Legal medicine 8.2 (2006): 110-120.

Out of environmental considerations **Table S2** belonging to this publication was not printed with this chapter of the thesis. The digital files can be obtained with the original publication at: https://doi.org/10.1002/humu.24068

Chapter 3

RMplex: an efficient method for analyzing 30 Y-STRs with high mutation rates

Arwin Ralf[1], Dion Zandstra[1], Natalie Weiler[2], Wilfred F.J. van Ijcken[3], Titia Sijen[2,4], Manfred Kayser[1]

[1] Department of Genetic Identification, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands
[2] Division of Biological Traces, Netherlands Forensic Institute, The Hague, the Netherlands
[3] Center for Biomics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands
[4] University of Amsterdam, Swammerdam Institute for Life Sciences, Amsterdam, the Netherlands

Published in *Forensic Science International: Genetics*, November 2021
Volume 55, 102595
doi: https://doi.org/10.1016/j.fsigen.2021.102595

Abstract

Y-chromosomal short tandem repeats (Y-STRs) with high mutation rates are recognized as valuable genetic markers for differentiating paternally related men, who typically cannot be separated with standard Y-STRs, and were shown to provide paternal lineage differentiation on a higher resolution level than standard Y-STRs. Both features make Y-STRs with high mutation rates relevant in criminal casework, particularly in sexual assault cases involving highly unbalanced male-female DNA mixtures that often fail autosomal forensic STR profiling for the male donor. Previously, the number of known Y-STRs with mutation rates higher than 10^{-2} per locus per generation termed rapidly mutating Y-STRs (RM Y-STRs) was limited to 13, which has recently been overcome by the discovery and characterization of 12 additional RM Y-STRs. Here, we present the development and validation of RMplex, an efficient genotyping system for analyzing 30 Y-STRs with high mutation rates, including all currently known RM Y-STRs, using multiplex PCR with capillary electrophoresis (CE) or massively parallel sequencing (MPS), overall targeting a total of 44 male-specific loci. If previously unavailable, repeat number assignations were provided based on newly generated MPS data. Validation tests based on the CE method demonstrated that the results were both repeatable and reproducible, full profiles were achieved with minimal input DNA of 250 pg for RMplex 1 and 100 pg for RMplex 2, and in the presence of inhibitors, or with a surplus of female DNA, the assays performed reasonably well. Application of RMplex to differentiate between paternally related men was exemplified in 32 males belonging to five different paternal pedigrees. Given further successful forensic validation testing, we envision the future application of RMplex in criminal cases where it is suspected, or cannot be excluded, that the crime scene trace originated from a male relatives of the suspect who is highlighted with standard Y-STR matching. Other applications of RMplex are in criminal cases without known suspects to differentiate between male relatives highlighted in familial searching based on standard Y-STR matching.

Introduction

Rapidly mutating Y-chromosomal short tandem repeats (RM Y-STRs) with mutation rates of 10^{-2} and higher (i.e., at least 1 mutation per locus every 100 generations), while standard Y-STRs have mutation rates of 10^{-3} and lower, were first described over a decade ago [1]. Since their discovery, RM Y-STRs have been demonstrated to effectively separate closely and distantly related males in paternal lineages [2, 3]. Differentiating male relatives is important in forensic Y-STR applications since male relatives typically show the same haplotype based on standard Y-STRs routinely applied in forensic DNA analysis, particularly closely related men such as fathers and their sons, brothers, or cousins. Therefore, after a match with standard forensic Y-STRs is obtained, it often remains unclear if the highlighted case suspect was indeed the crime scene sample donor, or if instead it was any of his paternal male relatives who share his standard Y-STR profile. Additionally, due to their increased diversity, RM Y-STRs were shown to outperform standard forensic Y-STR sets in male lineage identification, i.e., to differentiate between unrelated male individuals [4]. Both features make RM Y-STRs relevant in forensic casework where no autosomal STR profile is obtainable from the crime scene trace, as often observed in mixed male-female material typically available in sexual assault cases.

Over the recent years, the increased mutability of the 13 RM Y-STRs, which was initially described by Ballantyne et al. (2010) [1], has been confirmed in males from several populations such as from Italy [5], Serbia [6, 7], Turkey [8], Pakistan [2, 9], China [10-12], and Brazil. Despite occasional differences in observed mutation rates of the same Y-STR between different populations – which in part could have been the result of stochastic variations based on study sample size limitations – the increased potential of RM Y-STRs to discriminate between closely related males based on their high mutation rates was demonstrated in the different populations.

Because of their suitability to differentiate between both related and unrelated males, some RM Y-STRs were already included in the current versions of commercially available forensic Y-STR kits, such as Promega's Powerplex Y23 kit containing two RM Y-STRs [13], and Thermo Fisher Scientific's Yfiler Plus kit including six RM Y-STRs [14]. However, as these commercially available kits only include subsets of the 13 previously described RM Y-STRs, they do not maximize the differentiation capacity that is obtainable with RM Y-STRs. In contrast, several groups in academia had developed multiplex PCR

assays which targeted the full set of 13 RM Y-STRs [9, 10, 15, 16], highlighting the interest of the forensic genetic community in RM Y-STRs for casework applications.

A previous study showed that even with the full set of 13 RM Y-STRs, about three-quarters of father-son pairs could not be differentiated [2, 4]. Hence, to further increase the differentiation rate of male relatives, especially closely related ones, there was a need for more RM Y-STRs. The 13 previously described RM Y-STRs were discovered by means of a large empirical study, where we analyzed a large number of known Y-STRs (almost 190) in a large number of father-son pairs (almost 2000) to estimate their mutation rates [1]. Recently, we carried out a second search for Y-STRs with high mutation rates [17], in which we first applied a novel in silico approach to the Y-chromosome reference sequence for finding candidate Y-STRs with sequence features known to result in increased mutation rates, followed by their analysis in a large number of father-son pairs for empirical mutation rate estimation. In this recent study, we identified 12 novel RM Y-STRs with mutation rates $>10^{-2}$, and an additional ten Y-STRs with mutation rates between 5×10^{-3} and 1×10^{-2} that we had termed fast mutating (FM) Y-STRs to differentiate them from RM Y-STRs with mutation rates $>10^{-2}$ on one hand and standard Y-STRs with mutation rates $<10^{-3}$ on the other hand [17]. Combining the 12 novel with the 13 previous RM Y-STRs, and using the empirically established mutation rates to calculate the capacity of this set of 25 RM Y-STRs to theoretically discriminate paternal male relatives, we previously revealed that nearly half of father-son pairs are expected to be distinguishable with at least one mutation at any of these 25 RM Y-STRs [17]. If confirmed by empirical evidence from analyzing a large number of independent father-son pairs and other pairs of male relatives, this would provide a strong improvement in male relative differentiation compared to the previous set of 13 RM Y-STRs [2].

Here, we present the development of RMplex, the first multiplex genotyping method for the efficient analysis of 30 Y-STRs with high mutation rates, including all currently known 26 RM Y-STRs (with DYF403S1a and DYF403S1b considered as separate loci) together with 4 FM Y-STRs, targeting a total of 44 male-specific loci. Moreover, we provide preliminary empirical evidence for the suitability of RMplex for forensic applications, including typical elements of a developmental validation study such as reproducibility testing, repeatability testing, sensitivity testing, specificity testing, concordance testing, inhibitor testing, and testing of excess female DNA in male-female mixtures.

Materials and Methods

Y-STR ascertainment

A total of 30 Y-STRs with high mutation rates were included comprising all 26 currently known RM Y-STRs, i.e., the previously identified 14 RM Y-STRs from Ballantyne et al. 2010 with DYF403S1a and DYS403S1b considered as separate loci [1] and the recently described 12 RM Y-STRs from Ralf et al. 2020 [17], as well as four FM Y-STRs. These four FM Y-STRs DYF393S1, DYS442, DYS1005 and DYS1013, all had estimated mutation rates between 8 x 10^{-3} and 10^{-2}; hence, their classification as FM Y-STRs [17]. An overview of all 30 Y-STRs that are included in RMplex can be found in Table 1; the corresponding genomic locations relative to GRCh38 can be found in supplementary data Table S1.

Primer design

Reference sequences for all 30 Y-STRs, including 200 bp flanking regions, were obtained from the human reference genome GRCh38.p12 Y-chromosome sequence using Ensembl genome browser 95 [18]. Primer3 was used to design a maximum of five candidate primer pairs (cPPs) per each of the 30 Y-STRs [19]. Within Primer3 the "Targets" regions were specified as the repetitive regions and a maximum product size of 600 bp was defined. To confirm the specificity of the designed primer pairs, an *in silico* PCR was performed on all cPPs using BiSearch [20]; cPPs that produced nonspecific products were discarded. For each of the remaining cPPs, the allele range was estimated based on length of the *in silico* PCR products in the reference genome and on the allele ranges that were observed in previous studies [4, 17]. The cPPs were checked against each other for possible heterodimerization using the primer3-py API for Python (https://github.com/libnano/primer3-py). If for a given Y-STR no cPPs were found to be suitable after these steps, the flanking regions were extended and the "Included region" in Primer3 was modified to force the primer design deeper into the flanking regions.

After this initial primer selection step, sets of primer pair combinations were constructed using a graph-expanding greedy algorithm in a self-written Python program. This program accounted for heterodimerization by using a "compatibility matrix" for all cPPs of all Y-STRs and it avoided overlap of allele ranges of the different Y-STRs in the same channel by including 19 bp spacers. In brief, the algorithm first selects the STR with the lowest upper bound of its allele range and chooses that cPP. Secondly, the next compatible STR with the lowest upper bound of its allele range is selected. If the lower

bound of the allele range does not overlap with the allele range of the previous one, or with the spacer, and if the cPP is compatible with the previous one(s), it is assigned. This process is continued until the first color channel is filled and then repeated for the next color channel, again starting with the Y-STR with the lowest upper bound of allele ranges. When the first multiplex was designed in this way, the whole process was repeated for a second multiplex PCR assay containing the remaining Y-STRs. A similar approach was previously described by Shen et al. [21].

Multiplex PCR development

For the constructed primer pair combinations, unlabeled oligonucleotides were ordered (Integrated DNA Technologies) and tested in singleplex PCR using genomic DNA from two males; one female (as negative control); and a PCR blank as template. PCRs were performed with 30 cycles and an annealing temperature of 60˚C on a Veriti™ 96-Well Thermal Cycler (Thermo Fisher Scientific) using QIAGEN Multiplex PCR Plus Kit (Qiagen). The resulting PCR products were visualized on 2% (w/v) agarose gel including GelRed® Nucleic Acid Gel Stain (Biotium) at a 1X concentration.

Some loci showed nonspecific PCR products in the female sample or showed no product at all. Sometimes a different cPP could be selected in such cases, while in other cases a new primer design was inevitable. Changing the primer pairs of a single Y-STR could lead to a changed product size and thereby cause an overlap with the allele range of a second Y-STR in the same color channel. Moreover, a new primer pair could be incompatible with other primers in the same multiplex PCR. Hence, primer redesign was conditioned to allow for subsequent multiplexing. This process could lead to a highly different set of primers, even altering primer pairs that did perform well in the initial design. This process of primer redesigning was performed manually following the same criteria as described above, while using the original design from the algorithm as a backbone structure. The process was repeated until all Y-STRs could be successfully amplified, and the vast majority showed only specific products (remaining nonspecific products will be discussed later). The primer sequences that were used in the final multiplex assays are shown in Table 1.

All forward primers were ordered with one of four fluorescent labels as shown in Table 1. Singleplex PCRs were repeated and 1 µL of the amplification products was analyzed together with GeneScan™ 600 LIZ dye Size Standard v2.0 (Thermo Fisher Scientific) on an ABI 3500 Genetic Analyzer (Thermo Fisher Scientific), both POP-4™ and

POP-7™ Polymer for 3500/3500xL Genetic Analyzers (Thermo Fisher Scientific) were used. The genetic analyzer was equipped with a 36 cm 3500 Genetic Analyzer 8-Capillary Array (Thermo Fisher Scientific), while using the DS-33 Matrix Standard Kit (Thermo Fisher Scientific). The injection time was 15s at 1.2 kV, the fragments were run for 1400s at 15 kV. The resulting electropherograms were inspected in Genemapper IDX v1.5 (Thermo Fisher Scientific) panels and bin sets that were developed by testing a large number of male individuals and adding a new bin for every new variant that was detected.

Two multiplex PCR assays were designed for the 30 Y-STRs. Multiplex optimization was performed on a Veriti 96-Well Thermal Cycler (Thermo Fisher Scientific). The PCR mix consisted of 6.25 µL 2x Multiplex PCR Master Mix from the QIAGEN Multiplex PCR Plus Kit (Qiagen); 1.75 µL 5X AmpSolution™ Reagent (Promega); 2.5 µL pooled primer mix; and 2 µL of template DNA. Cycling conditions were initial denaturation of 10 minutes at 95 ˚C; followed by 29 cycles of: 95˚C for 30s, 64˚C for 30s and 72˚C for 30s; two final extension steps were included one at 72˚C for 10 min, followed by 45 min at 60˚C. Capillary electrophoresis was performed using the same conditions as described above. Peak heights were balanced by titrating the primer concentrations; the final primer concentrations can be found in Table 1.

Table 1: Y-STR included in RMplex and the RMplex multiplex assay details.

Y-STR locus	Mutation rate (x10^{-3}) [reference]	Dominant motif size	Forward primer sequence	Reverse primer sequence	RMplex assay	Fluorescent dye label	Size range (bp)	Number of copies	Concentration (µM)
DYF393S1	8.6 [1]	3	AAGCAGAGCCACA CAGACT	GTTTGCCTGTAAGTGG AGCC	1	FAM	158-204	1	0.15
DYS627	12.3 [1]	4	ACAGCGCAGGATT CCATCTA	TGCCTTTCATTCTCTCC TTCG	1	FAM	247-281	1	0.3
DYS570	12.4 [1]	4	GAGGAGATTAGG AGCACAGTGA*	TGCAAGGTGTGGGTG AAAATT	1	FAM	308-352	1	0.2*
DYS713	14.2 [17]	4	CTGGGTGTGCATT CGAGACT	GTTGCAGGGAGTGGA GATTG	1	FAM	410-450	1	1
DYS526b	12.5 [1]	4	GCCCTTGTTTCTAT AAGTGGTCA	GTTTGGGTTACTTCGC CAGA	1	FAM	473-523	1	0.4
DYF1000	35.9 [17]	3	CAGGGAGCTTCAG TGTGC	TGGCTCAGCTCACAGT AGAA	1	VIC	198-306	4	0.15
DYS518	18.4 [1]	4	TGGGCCAAGATCT CGTCATT	TCACATGTAGCACTCT GGGC	1	VIC	327-384	1	0.1
DYS1003	13.0 [17]	4	CAGTCAGCCAAGA TGCCAAA	GCAACACTTAAGAGAC GGCA	1	VIC	406-471	1	0.15
DYS1012	19.2 [17]	4	GCAAGACTCCATC TCCAAAAAG	CAAGCTTGGGTCCATT ATGA	1	NED	246-292	1	0.8
DYS1005	9.3 [17]	4	TGGATGGAAGTG GTACCTCTG	AGGTTGTGGTAATCTG AGATTGC	1	NED	356-403	1	0.2
DYS1010	14.2 [17]	4	CTACTCCAAAGGC TGCAGGA	CGCCCTCACACCCTTTC TTT	1	NED	451-494	1	1.2
DYS1007	15.5 [17]	4	GGTAAGATAATAT GGCACCGTGG	CCCTCCTTCCCTCCTTA TCTC	1	PET	243-305	1	0.3
DYR88	29.1 [17]	4	GCGAGACTCCATC TCAACA	CCACCCAAATCTCACGT	1	PET	343-378	2	1.2

Table 1 (continued)

DYF404S1	12.5 [1]	4	AGTACTTTGAGTT TCCCAGAAGG	AAGGAGCCCAGGATTG AGAG	1	PET	411-449	2	0.2
DYF387S1	15.9 [1]	4	GTCTCACTAGCTG GTCAGGG	GTCGTGGTGGTAAGTG CATT	1	PET	489-531	2	0.15
DYS1013	9.9 [17]	4	TCTGACTCCTTGA CTCCCAA	TGTCTCTCTGTCTGCCT GC	1	PET	552-588	1	1
DYS712	27.2 [17]	4	TTGAGCCCAGAAG TTCAAGAA	GGTACCTTGTATTTTCC ACAGGA	2	FAM	261-321	1	0.35
DYS711	26.6 [17]	3	TGGTGATTACATA TTGCACAGCC	GCTGTCATTGTATCTCT TCACTC	2	FAM	339-407	1	0.65
DYS626	12.2 [1]	4	AGCTGAGGAAGA GAATGGCG	GCAAATGTAAGTCTGT CTCTGGA	2	FAM	447-494	1	0.4
DYF399S1	77.3 [1]	4	TTGCATAGGTAGA GGGAGGC	GCTTAGGATTGGACCC AGGA	2	VIC	170-222	3	0.15
DYS449	12.2 [1]	4	GTCTCTCAAGCCT GTTCTATGA	TGGACAACAAGAGTAA GACAGAA	2	VIC	285-331	1	0.4
DYS724	46.4 [17]	4	TGATGGCTCATTG TAGTCCAC	AGCTGTTAACCTTCCCA AATTGT	2	VIC	352-418	2	0.2
DYS547	23.6 [1]	4	TCTGTTTCTGCATT GTTTCACTT	TGAGTGACAGAGCATA AACGTG	2	VIC	471-512	1	0.7
DYS576	14.3 [1]	4	TCTCAGCCAAGCA ACATAGC	TGGCAGTCTCATTTCCT GGA	2	NED	147-188	1	0.15
DYS612	14.5 [1]	3	GCAGAAAGGGCC TTAAGACA	CTTGACACTTGCCATG GGTA	2	NED	257-291	1	0.15
DYF1002	17.9 [17]	4	GCGAGGGGTAAG TAGTGGAA	ACATCACATCTCTCCTT CCTTCT	2	NED	317-363	2	0.45
DYF1001	52.0 [17]	4	GTTGTGGTGATCT GAGATTGCT	ACTGGATGGAAGTGGT ACCT	2	NED	437-508	3	0.6
DYF403S1a	31.0 [1]	4	GGYAACAGAGCA GGATTCCATCTA	ACATAGTTCAAAATTC ATGTGGATAATGA	2	PET	288-354	3	1

Table 1 (continued)

DYS442	9.8 [1]	4	CGGAGGAAAAGA AGTGATTGTAC	CCCCAAAGTGTGTTGC ATCA	2	PET	375-400	1	0.2
DYF403S1b	11.9 [1]	4	GGYAACAGAGCA GGATTCCATCTA	ACATAGTTTGAAATTC ATGTGGATAATGA	2	PET	417-471	1	1

*DYS570_alternative_forward primer: GAGGAGATTAGGARCACAGTGA; when using this primer the concentration should be doubled to 0.4 µM

Y-STR sequence analysis

Consistency in repeat number assignation between different kits is of high importance for practical applications. Here, we used the following approach to assign repeat numbers to the observed alleles. First, if a given Y-STR was already included in a commercial kit, i.e., DYS570, DYS576, DYS627, DYS518, DYF387S1, DYS449, DYS612, and DYS724 we adopted that same previously introduced repeat nomenclature here without any change to allow consistency. Secondly, for Y-STRs not included in commercial kits for which a repeat number assignation was already described in scientific literature (i.e., DYS711, DYS712, DYS713, DYF399S1, DYS547, DYS526b, DYS626, DYF403S1a, DYF403S1b, DYF404S1), we adopted that same repeat number nomenclature here without any change to allow consistency. Lastly, for Y-STRs that were not included in forensic kits and did not have a repeat number nomenclature in literature, we assigned a new repeat number nomenclature here based on de novo generated MPS data. To this end, all 30 Y-STRs included in the two developed multiplex assays were sequenced using MiSeq (Illumina) MPS technology.

For this, we amplified the 30 Y-STRs in three DNA samples using the newly developed two multiplex assays and under the same conditions as previously described, while using primers without fluorescent labels. The following commercially available DNA samples were used: AmpFℓSTR™ DNA Control 007 (Thermo Fisher Scientific), NA24385 / HG002 (Coriell Institute), and NA24631 / HG005 (Coriell Institute). Library preparation was performed using the Ovation Low Complexity Sequencing System (NuGEN). For each sample, 135 ng of the multiplex PCR was end repaired and followed by ligation of adaptors that are specifically designed to increase complexity in otherwise low complex amplicons. The adaptor ligated products were directly sequenced using paired-end 2x 300 bp v3 sequencing chemistry (Illumina) on a MiSeq system (Illumina). For some of the longer and more complex Y-STRs, the 2x 300 base pair paired-end sequencing reads were too short to fully span the amplicon including both primer sequences. Therefore, the software tools PEAR [22] and NGmerge [23] were used to, *in silico,* combine the read pairs into one longer read. An in-house software tool was used to analyze the resulting sequencing reads. This tool uses the primer sequences to group the obtained sequencing reads per each Y-STR. After grouping the reads, the in-house software tool counts the number of times that each unique sequence occurs. In most cases, it was obvious that the most frequently occurring sequence(s) described the true Y-STR allele and that other sequences were stutter alleles or contained sequencing errors. However, for some multi-copy Y-STRs the interpretation was somewhat more challenging, for example, there could be differences in the number of reads covering the different copies, multiple copies could

hold the same sequence, or a copy could be in the stutter position of another copy. As an additional confirmatory step, the length differences of the alleles between individuals, and between different alleles within individuals in the case of multi-copy Y-STRs, as obtained from the sequencing analysis were compared to the differences in allele length obtained by capillary electrophoresis. Additionally, for two of the DNA samples used, high-fidelity long read whole genome sequencing data generated with Pacbio's Sequel II was publicly available (accession numbers: PRJNA586863 for NA24385 and PRJNA540706 for NA24631); these data were used as an additional resource to help determining the correct sequences of the most challenging Y-STRs.

Y-STR repeat number assignation

To assign a repeat number to the generated sequences of Y-STRs for which a repeat nomenclature was previously unavailable, the recently published software tool STRNaming was used [24]. This software tool automatically converts sequence-based alleles to shortened allele names in bracketed format. To decide which elements of the bracketed format had to be included in the repeat number, the following criteria were used. All variable repeat stretches included in the amplicon were considered. If the Y-STR contained repeat stretches with different motif sizes, the longer, or shorter motifs were counted according to the dominant motif. For example, if a locus contained a tetranucleotide repeat stretch of 20 units and additionally held 13 repeats of a dinucleotide repeat stretch, the latter were also counted as six and a half (or rather 6.2) repeat units; hence, the repeat number assigned to the allele in this example would have been 26.2. Also, non-repetitive elements that varied between the obtained sequences were included in the nomenclature. Since the number of sequences obtained here was limited, and by no means sufficient to cover all variation that could be present in these Y-STRs, also repeat stretches not found to be variable in these sequencing data were included in the repeat nomenclature as long as their length was at least 12 nucleotides, i.e., 2, 3, 3, 4, and 6 repeat units for hexa, penta, tetra, tri, and dinucleotide repeats, respectively. Although by applying this system, repeat numbers were assigned to all Y-STRs included in the multiplex assay, for all Y-STRs that are included in commercial assays, or where a repeat nomenclature was previously published (i.e, [4, 25]), we used that previous nomenclature here.

Validation tests

All experiments for the validation tests were performed using the PCR conditions as described in 2.3, using a GeneAmp™ PCR System 9700 (Thermo Fisher Scientific). The

resulting amplicons were separated using ABI3500 Genetic Analyzers (Thermo Fisher Scientific) using a 36 cm capillary array and using the DS-33 Matrix Standard for spectral calibration and GeneScan 600 LIZ Dye Size Standard (Thermo Fisher Scientific) using POP4 (Thermo Fisher Scientific). The electropherograms were analyzed using Genemapper IDX v1.5 (Thermo Fisher Scientific) with a minimum of 150 RFUs for peak detection. Global Minus Stutter Ratios were set to 0.8 for trinucleotide repeats and to 0.6 for tetranucleotide repeats.

Repeatability

To test the repeatability of RMplex, DNA sample NA24385 was amplified with the two developed multiplex PCR assays in twelve individual PCRs. CE was performed in 24 different runs on the very same ABI 3500 Genetic Analyzer (Thermo Fisher Scientific) machine. Genotype concordance, fragment length variation and peak intensities were compared between the different repetitions.

Reproducibility

To test the reproducibility of RMplex, a total of 20 male samples from the ECACC Ethnic Diversity DNA Panel (Sigma-Aldrich) were analyzed by two laboratories. The first laboratory used a Veriti™ 96-Well Thermal Cycler (Thermo Fisher Scientific) for amplification and analyzed the data using Genemapper IDX v1.5 (Thermo Fisher Scientific). The second laboratory used a GeneAmp™ PCR System 9700 (Thermo Fisher Scientific) for amplification and GeneMarker HID v2.9.8 for the analysis. Both laboratories applied an ABI 3500 Genetic Analyzer (Thermo Fisher Scientific) with a 36 cm array and POP4 for amplicon separation. The genotyping data obtained by both laboratories were compared for concordance. Moreover, RMplex was used to genotype a total of six case samples, which had previously been typed with the ForenSeq™ DNA Signature Prep Kit (Verogen) that overlaps with RMplex in four Y-STRs, and these data were used for additional concordance testing. The six cases were from the years 1991, 1995, 2005, 2006, 2014 and 2019, three were sexual assaults and three were other types of criminal cases. For the six samples, the total human DNA concentration per sample ranged from 0.46 to 22.42 ng/µl and the male DNA concentration ranged between 0.29-16.20 ng/µL, which allowed the generation of informative autosomal DNA profiles.

Sensitivity, Specificity, and Robustness

To test for the sensitivity of RMplex, a dilution series was prepared using a single DNA sample with concentrations 50, 100, 250, 500, 1000, and 2000 pg/µL; 1 µL of each

concentration was used as input for the PCRs in triplicate. Human and male specificity was tested via the analysis of 10 ng of DNA from a human female, a mouse, a cow, a dog, a cat, a chicken, a pig, a rat and a chimpanzee with both multiplex assays, the sex of all animals was unknown, except for the male chimpanzee. The effect of two inhibitors: humic acid and tannic acid on the performance of the multiplex assays was tested. For each inhibitor, a high and a low concentration was tested. For humic acid the low concentration was 25 ng/µL and the high concentration was 75 ng/µL; while for tannic acid these were 5 ng/µL and 20 ng/µL, respectively. The effect of the inhibitors was tested in duplicate using different gDNA input amounts ranging from 100 pg to 750 pg.

Male-female mixtures

To test the performance of RMplex in the presence of a surplus of female DNA, male-female DNA mixtures with seven different male:female input amount ratios were prepared, i.e., 1:1, 1:10, 1:20, 1:50, 1:100, 1:200, and 1:500. Two different amounts of male input DNA were used, to prepare mixtures with the respective ratios. All mixtures were genotyped in triplicate.

Degraded DNA

To test the performance of RMplex with low quality DNA we prepared a series with increasingly fragmented DNA using a Covaris S220 Focussed-ultrasonicator (Covaris). DNA fragmentations were carried out on a genomic DNA sample of 10 ng/µL in volumes of 15 µL with a Peak Incident Power of 18 W, a Duty Factor of 20%, 50 cycles per burst and varying treatment times ranging from 10s to 500s. The fragmented DNA was analyzed using a 2100 Bioanalyzer System (Agilent) using DNA 1000 chemistry (Agilent). The fragmented DNA was diluted to a concentration of 1 ng/µL and 1 µL of those dilutions were genotyped using the RMplex in triplicate.

Male relative differentiation

To get a first impression of the capacity of the new method to differentiate between male relatives within paternal pedigrees, the paternally related males of a total of five CEPH Reference Families (Coriell Institute) were genotyped with RMplex: CEPH/French Pedigree 66; CEPH/Utah Pedigree 1362; CEPH/Utah Pedigree 1423; CEPH/Utah Pedigree 1454; and CEPH/Utah Pedigree 1463. Together, these five families cover 27 paternally related male relative pairs separated by one meiosis and 66 pairs separated by two meioses (40 brother pairs and 22 grandfather-grandson pairs). From these data, preliminary relative differentiation rates were calculated per each group of male relatives.

Results & Discussion

RMplex method development

We successfully developed RMplex, a multiplex genotyping method for efficient analysis of 30 Y-STRs with high mutation rates in two multiplex assays targeting a total number of 44 male-specific loci. The number of loci is higher than the number of Y-STRs because of the presence of several multi-copy Y-STRs. RMplex 1 includes 16 Y-STRs of which two have a dominant trinucleotide repeat motif and 14 a dominant tetranucleotide repeat motif, while 12 Y-STRs are single-copy Y-STRs and 4 are multi-copy Y-STRs; altogether a total of 22 male-specific loci are targeted with RMplex1. RMplex 2 targets 14 Y-STRs of which two have a dominant trinucleotide repeat motif and 12 a dominant tetranucleotide repeat motif, while 9 are single-copy and 5 are multi-copy Y-STRs, with a total of 22 male-specific loci targeted. The distribution of the 30 Y-STRs over the two CE-based multiplex assays and over the four fluorescent dye channels is visualized in Figure 1. Electropherograms showing examples of both multiplex assays when applied to the AmpFℓSTR DNA Control 007 DNA are provided in the supplementary material in Figure S1 and Figure S2.

We noticed that the graph-expanding greedy algorithm had some limitations, i.e., the initial iterations resulted in incompatible leftover cPPs and a minimum of three multiplex assays would have been required to combine all Y-STRs following those iterations. We found that the length of the spacers had a large impact on the final success of the algorithm, where finally with a 19 bp spacer we reached the optimal result. Ideally, an algorithm that would test all possible combinations of cPPs might have been more effective. However, the computational cost of such an algorithm would be orders of magnitude larger than the graph-expanding greedy algorithm which eventually also led to a positive outcome. Given the length distributions of the alleles observed in previous studies [4, 17], none of the alleles between Y-STRs in the same fluorescent dye channel are expected to overlap. However, some of the Y-STRs were predominantly tested in males of European descent thus far; therefore, alleles with lengths well outside the currently known range may exist in populations other than Europeans, which needs to be monitored carefully in future studies. Overall, during the multiplex optimization, the two RMplex assays worked well and delivered complete and reproducible genotypes for all 30 Y-STRs included and for all 44 male-specific loci that were targeted.

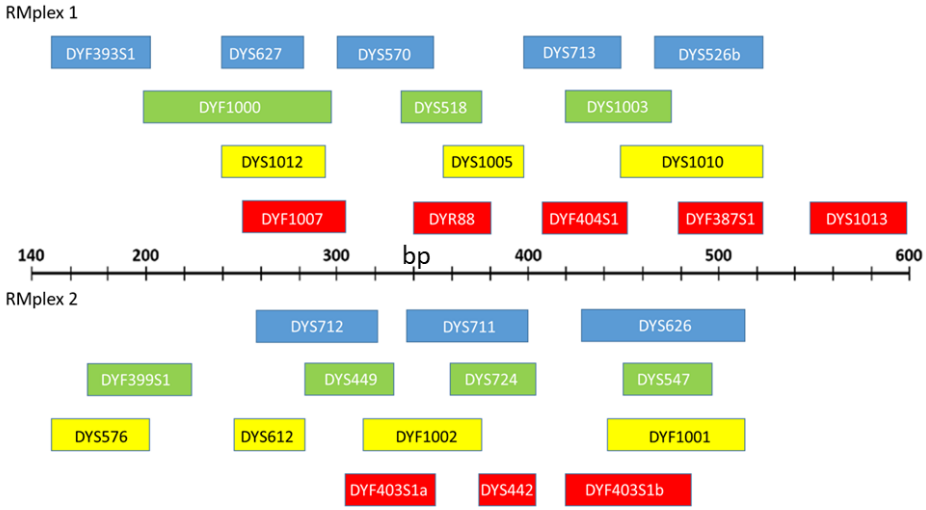


Figure 1: The position of the 30 Y-STRs included in RMplex 1 and RMplex2 over currently known allelic fragment length space and the four fluorescent dye channels used for CE analysis. The colors represent the color of the fluorescent dyes that were used.

General observations

Some of the 30 Y-STRs included in these newly developed assays behave somewhat differently compared to Y-STRs included in commercial Y-STR kits. In the following, we provide a short description of such Y-STRs combined with a discussion on how this may impact the result interpretation.

Stutter

Some of the 30 Y-STRs exhibit high stutter peaks, which in part may be explained by the sequence features that cause these Y-STRs to mutate frequently *in vivo,* e.g., a large number of uninterrupted repeats [1, 17]. The same feature has been shown to make STRs more prone to slippage artifact during PCR *in vitro* [26]. In consequence, one may argue that observing increased stutter in Y-STRs with increased mutation rates might not be unexpected. High stutter alleles were particularly observed in trinucleotide repeats. The most remarkable example is DYS711 as shown in Figure 2a, which is a single-copy Y-STR (based on the reference sequence); hence, all peaks other than the highest can be considered as stutter alleles. The -1 stutter allele shows stutter ratios of above 70%; moreover, also -2, -3 and +1 stutter alleles are seen. Another example, as shown in Figure 2b, is the trinucleotide multi-copy Y-STR DYF1000. Although in general the stutter ratios are smaller for DYF1000 compared to DYS711, ratios of above 50% were not commonly

observed. Because this is a multi-copy Y-STR, if two true alleles are adjacent to each other, these relatively high stutter peaks may impact each other, which can lead to stutter alleles and true alleles to be of almost equal peak height, as shown in Figure 2b. Nevertheless, when taking these high stutter ratios into account, it generally is quite straightforward to call the alleles correctly, especially when assigning high stutter ratio thresholds to such Y-STRs in the analysis software. In our experience, values of 0.8 for trinucleotide repeats and 0.6 for tetranucleotide repeats are suitable for analyzing the Y-STRs correctly in an automated manner. However, in cases involving mixtures with multiple male contributors, these high stutter ratios would likely complicate unambiguous interpretation of the results; more so, interpretation of multi-copy Y-STRs could become increasingly challenging. This aspect was not tested in this study, and it would be recommended to perform a comprehensive empirical study on male-male mixtures prior to applying this method in such scenarios.

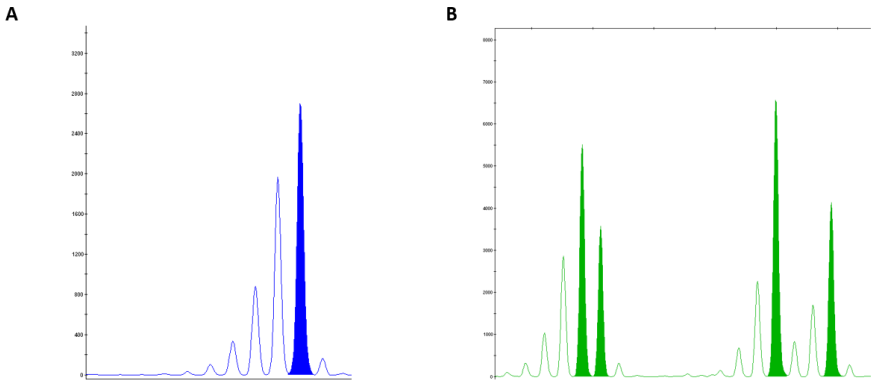


Figure 2: Examples of two trinucleotide RM Y-STR with remarkably high stutter artifacts obtained with RMplex: DYS711 (A) and DYF1000 (B), the true alleles are highlighted.

Multi-copy Y-STRs

Although multi-copy Y-STRs have been included in commercial Y-STR kits occasionally (i.e., DYS385, and more recently DYF387S1), they play a more dominant role in the RMplex assays. In multi-copy Y-STRs, mutations can independently occur at each of the copies as they represent separate male-specific Y-chromosome locations, therefore undergoing independent recombination slippage creating repeat mutations. As long as these multiple copies cannot be analyzed separately with a suitable genotyping assay, combining them in the mutation rate estimate consequently leads to an increase in the overall mutation rate for such multi-copy Y-STR. Therefore, multi-copy Y-STRs are enriched in sets of Y-STRs with high mutation rates. Commercial assays have been limited to multi-copy Y-STRs that,

typically, only have two copies; although, occasionally observed additional duplication events can lead to more copies and alleles for those Y-STRs too [27, 28]. Here, we included multi-copy Y-STRs containing up to four copies in the general population. Additional multiplication events at such Y-STRs could lead to the presence of an even large number of alleles in such a multi-copy Y-STR.

Interpreting peak height differences in multi-copy Y-STRs, especially those containing more than two copies, can be challenging. An example is presented in the supplementary material; Figure S3, which shows the genotypes from two related males for DYF1001, which typically has three male-specific copies. In the upper electropherograms there is a clear difference in peak height between the allele A (8084 RFUs) and B (14539 RFUs). A valid interpretation could therefore be that the genotype is A/B/B. However, the possibility of a locus deletion having removed one of the three copies, leaving this man with only two copies, and the true genotype consequently being A/B cannot be totally ignored, as peak height differences between different copies can also be the result of stochastic PCR effects. In this case, a paternal relative of this individual was genotyped who displays a different genotype (bottom electropherograms of Figure S3). Here, we can see that three copies are present (A/B/C). This suggests that the true genotype for the individual in the upper electropherogram likely is A/B/B (instead of A/B) and a mutation occurred between these two relatives, so that one of the B-alleles became the C-allele in the relative shown in the bottom electropherogram (or alternatively the C-allele could have become one of the B-alleles, depending on which is the ancestral genotype). In principle, how to interpret such multi-copy Y-STRs would only be an issue when storing the data in a database. In that scenario, it would require a convention to avoid subjectivity in the calling of such genotypes. Perhaps not taking peak height into account would then be the most conservative, safe, and preferred approach. However, direct comparisons to find allelic differences between of two related potential suspects is much less affected by this issue. Notably, this is currently the main forensic application of RM Y-STRs.

Microvariants

Many RM Y-STRs have complex repeat structures, and in some cases, they include repeat stretches with repetitive motifs of different sequence and lengths. As a result, microvariants are observed more commonly at RM Y-STRs than at standard Y-STRs that typically have a less complex repetitive structure. For example, Supplementary Figure S4 shows alleles observed at DYS1005, which on top of multiple tetranucleotide repeat stretches also contain a variable pentanucleotide repeat stretch. Variations in the pentanucleotide repeat result in microvariants as shown. Additionally, there are also

multi-copy Y-STRs where different copies can contain a microvariant caused by an insertion/deletion polymorphisms (indel) e.g., DYF399S1, or multi-copy Y-STRs with apparent microvariants because of more pronounced sequence differences between the amplified alleles (i.e., DYF403S1a and DYF1000). The complexity of such Y-STRs could be especially valuable if they were sequenced instead of analyzed via CE. However, when using CE, sizing accuracy and resolving power are critical. The latter is especially important in multi-copy Y-STRs when two alleles are present with only one nucleotide size difference due to the presence of a microvariant.

Locus-specific observations

We noticed that in some DNA samples, DYS570 was amplified with relatively low intensity, or dropped out completely, despite the other Y-STRs performing well. A closer inspection revealed a flaw in the primer design for DYS570: the forward primer that was used contained a SNP: rs9786374 (also known as L537, or PF1507), which phylogenetically is located at the root of Y-haplogroup E (according to the ISOGG Y-DNA Haplogroup Tree). This would mean that, if unmodified, our method would perform poorly for DYS570 in DNA samples belonging to haplogroup E. Given the known worldwide distribution of this haplogroup, this would affect the genotyping of many individuals from the African continent, or Western Asia, but also some Europeans. In Table 1 we present a modified primer with a degenerate base to overcome this amplification issue. By comparing the performance of RMplex 1 with the original forward primer and with the modified forward primer in a haplogroup E derived and a non-haplogroup E derived individual, we showed that indeed the modified primer overcame the dropout that occurred with the original primer in the DNA samples from the haplogroup E individual, while not affecting the performance in the non-haplogroup E DNA sample (Figure S5). Hence, we strongly recommended to use the modified primer and adjusted primer concentration as described in Table 1. However, during some of the validation tests in this study, the original DYS570 primer was used, often in combination with a male control DNA sample that belongs to haplogroup E, which has likely affected some of our outcomes as described in section 3.3. Moreover, it has to be noted that the multiplex assays have yet to be tested on a comprehensive population dataset that would be representative for the majority of the Y-chromosomal haplogroups present in modern humans. Until such additional testing has been performed, we cannot exclude the possibility that other loci dropping out in males belonging to other haplogroups. It is very likely that many other, yet unobserved, population-, or haplogroup-specific variants may exist that will show up in comprehensive population studies to be carried out in the future.

For all Y-STRs except one (DYS1012), primers could be designed without apparent by-products not being specific to males. Although according to our *in silico* analysis, the DYS1012 primers would be expected to work in a specific manner, in our experiments by-products were formed. This indicates that the region that is responsible for these DYS1012 by-products may not be represented in the current reference genome (GRCh38) used here. These products were observed in both male and female DNA samples; hence it is likely that they originate from autosomal loci or the X-chromosome. Furthermore, they are variable in size, making it likely that they contain repeats too. There appeared to be two groups of by-products, as exemplified in Figure 3, one with sizes ranging from ~80-120 bp and another ranging from ~170-210 bp. These groups do not overlap with any Y-STRs in the same channel as the shortest Y-STR is DYS1012 with a size range from ~230-310 bp (Figure 3); thus, we generally do not expect practical problems in correct allele calling. However, cases have been observed where by-products from the second group were significantly longer and thereby did fall in the size range of DYS1012. In such cases, accurate and unambiguous genotyping of this Y-STR is hampered, and it would therefore be advisable to exclude this single-copy Y-STR from analysis when multiple alleles are detected.

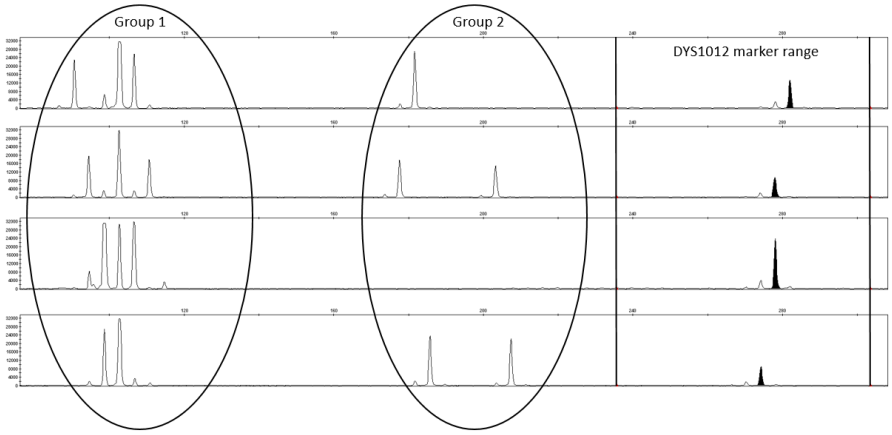


Figure 3: Examples of non-specific by-products generated by the primers used to amplify DYS1012.

Repeat number nomenclature

Assigning repeat numbers to (Y-)STR allele genotype calls is important as data generated in different laboratories need a convention to make them comparable and compatible. However, despite recommendations being in place, such as from ISFG [29], assigning a single repeat number to an allele that covers all sequence variability is challenging, if possible at all. Here we sequenced, using MPS, a total of three DNA samples to capture some of the variability present at the 30 Y-STRs included in RMplex; combined with the

GRCh38 reference sequence, we used four sequences for single-copy Y-STRs and more for multi-copy Y-STRs. Table 2 shows the structures and variations that were observed within these sequences using STRNaming [24]. The repeat structures in bold indicate those parts of the sequences that were used to assign the repeat number to an allele. For the Y-STRs that already had a repeat number nomenclature assigned because they were included in a commercial kit, or had been described in scientific literature, we adopted the previously described repeat nomenclature as shown in Table 3. Table 3 also shows the genotypes of two DNA samples commonly used as positive controls in forensic genotyping assays and of two commonly used and commercially available reference DNA samples. For the three DNA samples that were *de novo* sequenced as well as the GRCh38 reference sequence, the description of the repeat structures and the rationale behind the allele nomenclature can be found in Supplementary Tables S2-5.

Table 2: The repeat structures and observed sequence variability, the structures in bold were used for repeat number assignment for the new markers. The variability was based on three sequenced samples and the GRCh38 reference sequence.

Marker name	Repeat structure
DYF393S1	**GAA[4]AAG[22-25]CAG[0-1]AAG[0-1]**
DYS627	GA[3]CACA[1]GGGA[1]**GAAA[3]**GGAAAA[1]**AAGA[3]**GA[4]**GAAA[19-22]**AAGA[1]GAAA[1]GA[3]GGGA[2]
DYS570	**TTTC[17-21]**
DYS713	**TTTC[16-19]**TCTT[2]TCTG[1]TCTT[1]T[1]TTTC[2]**TCTT[12-14]TTTC[5]**[]TTAT[1]**TATT[4]**
DYS526b	CTTC[1]**CCTC[3]**TTTC[1]TTCC[1]TTTC[2]**TTCC[1-2]TTTC[11-13]CTTC[9]**TTTC[1]CTTC[1]TTTC[1]CTTC[1]ATTT[1]CTTC[1]TTTC[2][]**TCCT[12-14]**
DYF1000_1	**AAG[9]**CAG[1]**AAG[22-49]AGG[0-1]AAG[0-1]AGG[0-1]AAG[0-1]AGG[0-1]AAG[0-26]**
DYF1000_2*	**AAG[5]AAGGAG[6-7]**AAG[3]GAG[3]GA[1]**GAA[5-7]**GGA[2]**AGA[5]**AGG[1]AGA[1]GAA[2]G[1]**GAA[4]**GAT[1]GAA[2]**AGA[23-28]**
DYS518	GAAA[2]GGAA[1]**GAAA[15-16]**GGGA[1]**GAAA[4]**GGAA[1]GA[1]**GAAA[11-14]**GGAA[1]GA[1]GGGA[2]AGAA[1]GAAA[1]AGAA[2]**GGAA[3]**
DYS1003	AAAG[1]AAGA[2]AGGA[1]**AAGA[3]**AAGG[1]AACG[1]**AAGG[10-11]AAGA[3]GAAA[3]**AAAG[1]**AGGG[3]**AAGG[1]**AGGA[3]**[]GA[4]ACGG[1]**AAGA[17-19]**AAGG[1]TAGA[1]**AAGA[9-15]GAAA[5]AAGA[3]**
DYS1012	**AAGA[13-15]**AAGG[1]AAAG[1]AA[1]AAGA[2]**AAAG[3]**AA[1]**AAGA[5]AAAG[3]**AAGG[1]AAAG[1]AAGG[1]AAAG[1]AAGG[1]AAAG[1]AAGG[1]AAAG[1]AAGG[2]**AAAG[15-18]**
DYS1005	**TTCCT[12-13]TCCC[6-7]**TTCT[2]**CTTT[5]CCTT[13-14]CTTT[10-12]**CTT[1]**CCTT[4]**CTTT[2]
DYS1010	**AAGA[17]GA[12-15]AAGA[11-14]**AAA[1]AAGA[1]GAGG[2]GA[3]
DYS1007	**AAGA[13-17]GAAA[2-5]**GAGA[1]**AAGA[4-5]GAAA[14]**[]GAAA[1]GA[5]CAGC[1]AAGA[1]GAAA[1]GA[5]
DYR88	**AAAG[15-22]**
DYF404S1	**TTTC[14-16]**TCTC[1]TTTC[1]
DYF387S1	A[7]**GAA[4]**AGAG[1]A[7]GAAA[2]GGTA[1]**GGAA[4]**GAAA[2]GGAA[1]GAAA[2]**GGAA[9-10]GAAA[14-18]**AT[1]A[8]
DYS1013	**AGAT[11-12]**TGAT[1]AGAT[1][]AGAT[2]AGAA[1]**AGAT[6]**[]**GATA[10-11]**[]**AGAT[10-12]AGGC[3]**
DYS712	**AGAT[12-18]AGAC[5-7]**
DYS711	**TTC[4]**T[1]**TTC[29-33]**TCC[3]TTC[1][]**CTT[24-26]**CTC[2]CTT[3]
DYS626	AAAG[1]A[9]GCAG[1]AAAG[1]GAAG[1]A[5]**GAAA[17-19]**[]**GAAG[3]AAAG[3]**AGACAG[1]**AAAG[4]**A[6]GAAA[2]GAAG[1]**GAAA[3]**
DYF399S1	**AAAG[3]**AAAAAGA[1]**AAAGG[0-1]AAAG[16-20]AAAA[0-1]_+2.1->AAAA_+19A>G[0-1]**
DYS449	**TCTT[11-15]**TC[4]TCC[2]TCTT[2]TCCT[1]**TCTT[3]**TTCC[1]TCTT[1]TCCT[1]**TCTT[14-19]**
DYS724	CTTT[2]**TTTC[4]**TCTC[1]TTTC[2]TC[5]**CCTT[13-19]**T[1]CCTT[1]**CTTT[19-21]**
DYS547	**CTTC[3]TCCT[0-1]**TTTC[1]**CTTC[11]**CTTT[1]**CTTC[5]**TTTC[1]CTCC[2][]CTCC[1]CTTC[1]CTCC[1]TTCC[1]CTCC[2]TTTC[1]T[5]**CTTT[9-15]CTAT[0-1]CTTT[0-5]**CCTTA[1]TTTC[1]**TTCC[4]**TTTC[1]**TCTT[9-13]**TTCT[2]CTCT[1]**TTTC[3]**T[4]CTTT[2]T[6]
DYS576	**AAAG[17-19]**
DYS612	**CCT[5]**CTT[1]**TCT[4]**CCT[1]**TCT[25-27]**
DYF1002*	**GA[7]GAAA[4]**TAGA[1]**AAGA[11-17]AAGG[0-1]GAAA[4-5]**GGAA[1]**AGAA[10-12]GAAA[0-2]A[0-1]**CAGAGA[1]**AAGA[4]GAAA[10-14]**AAGA[1]A[1]AAGA[2]GGAA[1]**GAAA[4]**AAGA[1]GAAA[2]**GGAG[3]**GGAA[2]AGAA[1]GGAA[1]GGAG[1]
DYS1001*	**GAAG[9-15]**[]**AAGA[10-16]GAAA[0-1]GA[2-3]AAGA[4]**GAAA[1]**GA[2-4]**GAAA[1]GA[3]AAATGAAAGAAAAAGAA[1]**GGAA[4-6]GAAA[11-17]TAAA[0-1]**AGAA[1]**AAGA[4]AGGA[3]**AAGA[1]**GAAA[3]GGAA[0-1]GAAA[0-1]GGAA[10-15]**
DYF403S1a_1	TTCT[2]TTCC[1]**TTCT[10-15]**ATC[1]**TTCT[13-17]**
DYF403S1a_2	TTTC[2]ATTC[1]TCTT[2]TTTC[1]TTTT[1]**TCTT[14-17]TTTC[4]**
DYS442	**GATA[12]GACA[3]**GATA[2]
DYF403S1b	**GAAA[4]AAGA[11-15]AGGA[9-12]AAGA[9-12]GAAA[10-12]**AAGA[2]GAAC[1]GAAA[2]

* The nomenclature obtained from STRNaming was modified for at least one of the copies to increase the compatibility between the different copies

Table 3: Genotypes for two commonly used forensic control DNA samples and two commonly used and commercially available reference samples.

Marker	Control DNA 007	2800M Control DNA	NA24385/HG002	NA24631/HG005	Nomenclature adopted from
DYF393S1	31	25	28	28	This study
DYS627	21	22	21	24	Yfiler™ Plus
DYS570	17	17	18	21	PowerPlex® Y23 & Yfiler™ Plus
DYS713	42	45	44	42	Zhang et al. 2012 [25]
DYS526b	36	36	35	35	Ballantyne et al. 2014 [4]
DYF1000	57/58/70.2*/72.2*	55/62/70.2*/71.2*	59/60/69.2*/72.2*	58/71.2*/74.2*	This study
DYS518	37	36	39	35	Yfiler™ Plus
DYS1003	67	64	60	67	This study
DYS1012	43	39	39	43	This study
DYS1005	57.1	57.1	55.1	55.1	This study
DYS1010	35	31.2	37.2	37	This study
DYS1007	37	41	34	41	This study
DYR88	20/22	16	15/18	17	This study
DYF404S1	14/16	13/16	16	14	Ballantyne et al. 2014 [4]
DYF387S1	35/37	37/38	38/40	38	Yfiler™ Plus
DYS1013	40	41	43	42	This study
DYS712	19	19	18	25	Zhang et al. 2012 [25]
DYS711	66	56	61	63	Zhang et al. 2012 [25]
DYS626	30	29	31	32	Ballantyne et al. 2014 [4]
DYF399S1	24/26.1	24/25.1/26.1	20/21/25.1	22.1/23.1	Ballantyne et al. 2014 [4]
DYS449	30	34	28	32	Yfiler™ Plus
DYS724	35/37	36/38	32/37	38	FamilyTreeDNA
DYS547	49	45	46	49	Ballantyne et al. 2014 [4]
DYS576	19	18	18	17	PowerPlex® Y23 & Yfiler™ Plus
DYS612	31	29	31	32	ForenSeq DNA - Signature Prep
DYF1002	55.3/62.2	54.2/57.2	60.3/61.3	55.3/62.3	This study
DYS1001	71.2/72/77	75/75.2/78.2	75/79.2	75/76.2	This study
DYF403S1a	11/13/18.1*	11/14/17.1*	13/14/15.1*	12/15/16.1*	Ballantyne et al. 2014 [4]
DYS442	15	14	15	15	This study
DYF403S1b	51	46	50	42	Ballantyne et al. 2014 [4]

* These particular copies have a very different sequence from the other copies; because they are analyzed together, the allele nomenclature is based on the other two copies, also see Table S2-S5.

Validation testing

In our preliminary validation testing of the RMplex system, we included typical elements of forensic developmental validation studies.

Repeatability

A single DNA sample was analyzed in 12 separate PCR and CE runs, Figure 4 shows the standard deviation of the allele sizes of each Y-STR in RMplex 1 (A) and RMplex 2 (B); and the mean peak intensities per Y-STR in RMplex 1 (C) and RMplex 2 (D). Overall, both the allele sizing and the peak intensities were relatively constant between the different technical repetitions of the same sample.

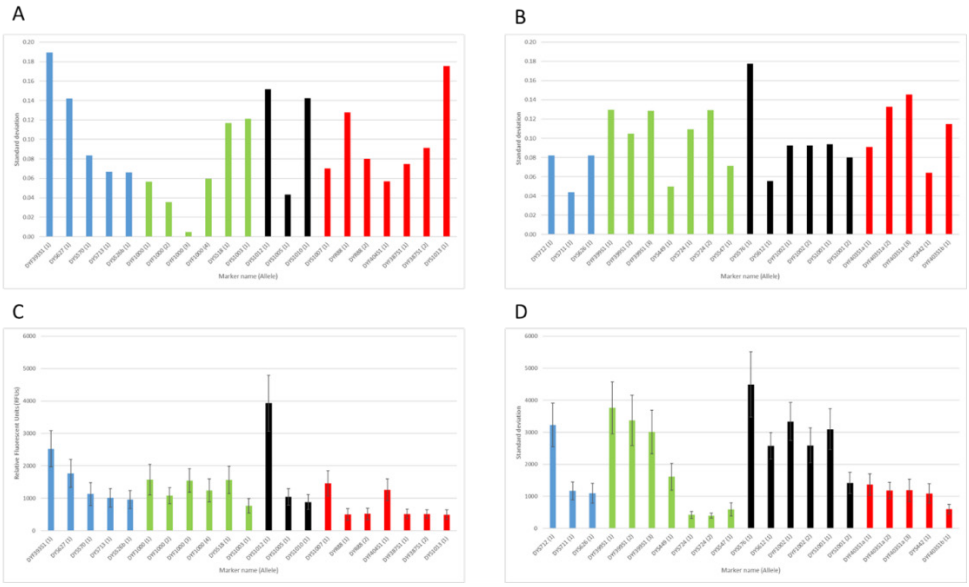


Figure 4: Length and intensity variations for each of the 30 Y-STR based on 12 genotyping repetitions of DNA sample NA24385; the standard deviations of the measured fragment in RMplex 1 (A) and RMplex 2 (B); and the mean peak intensities for each of the 30 Y-STR, where the error bars represent the 95% confidence intervals using the t-distribution, in RMplex 1 (C) and RMplex 2 (D). The colors represent the color of the fluorescent dyes that were used.

Reproducibility

To test the reproducibility of RMplex, 20 male DNA samples were analyzed independently by two different laboratories. A total of 803 alleles were detected among these 20 males. The two laboratories reported the same results for a total of 795 alleles (99%). In total,

there were eight discrepancies between the two laboratories: seven of these involved a single base pair (e.g., 34.1 vs. 34.2) and one a two base pair difference. Half of the discordant results were found at DYS1010, two at DYS1012 and one at DYF1001 and DYS1003, respectively. The fact that each of the discordant results could be classified as a different interpretation of microvariants, instead of repeat discordance, underlines the need for accurate sizing of the alleles when using CE. A way to help minimize such effects would be the development of a high-quality allelic ladder including all alleles for each of the Y-STRs included in both multiplex assays. However, the development of such ladders would require population studies to identify all relevant alleles. Moreover, developing such a ladder requires a considerable amount of effort and skill, given the high diversity of the Y-STRs, the high number of microvariants and the presence of multi-copy Y-STRs, and may therefore best be left to commercial parties. Nevertheless, the overall concordance between the genotypes obtained by both laboratories was high. Also, in a direct comparison of two or more haplotypes from male relatives established by the same laboratory, as in the typical forensic case work scenario of Y-STR application, this type of inter-laboratory difference would be of little importance.

Additionally, six forensic case samples that had previously been analyzed with the ForenSeq DNA Signature Prep Kit were analyzed with RMplex. This comparison was especially interesting because of the different technologies that were used to analyze the samples i.e., massively parallel sequencing for ForenSeq and CE for RMplex. In total, there were four overlapping Y-STR Y-STRs between ForenSeq and RMplex: DYS570, DYF387S1, DYS576 and DYS612, which could therefore be used for concordance purposes. In all cases when a genotype was obtained by both methods the result was concordant. However, DYS570 dropped out in one sample using RMplex; a haplogroup prediction based on the haplotypes obtained by the Forenseq kit and using the NEVGEN Y-DNA Haplogroup Predictor revealed that this individual belonged to haplogroup E, which could explain the drop out (see 3.2.4). Furthermore, in one sample two Y-STRs were not detected using ForenSeq, while RMplex did provide genotypes for both these Y-STRs.

Sensitivity

The sensitivity of RMplex was studied by genotyping a single DNA sample with input amounts varying between 50 pg and 2000 pg analyzed in triplicate. Within RMplex 1, allelic dropouts were detected at 50 pg DNA with a total of five dropouts i.e., two times one of the alleles at DYF404S1, once at each of DYS713, DYS570, and DYR88. In the 100 pg input samples, a total of three dropouts were observed, which were all at DYS570. Within RMplex 2, dropouts were only observed in the 50 pg input samples with six dropouts in

total: three times at DYS547, one time at each of DYF1001, DYS711 and DYS724. In the samples with 250 pg and with higher input DNA amounts, full profiles were obtained for all Y-STRs in all cases. Based on these, albeit preliminary, data it is recommended to use a minimal amount of 250 pg DNA as input for both multiplex assays of the RMplex system. Moreover, when using the redesigned DYS570 primers (see above and Table 1), the DYS570 drop-outs seen with the original primers at 100 pg would likely disappear, in which case the sensitivity threshold of the method based on both multiplex assays would be reduced to 100 pg. More work is needed to determine the final sensitivity threshold of these two assays.

Specificity

Overall, RMplex seems to be quite specific for humans and for human males. Regarding male-specificity testing in humans, the female human DNA sample only showed very few, low intensity peaks within the allele range of the 30 Y-STRs. Within RMplex 1, the most notable were peaks sized at ~300 bp in the yellow channel and in the red channel (Figure S6A). Within RMplex 2, no noteworthy peaks were detected (Figure S6B). Regarding human specificity testing, little cross-amplification with DNA from animal species was detected. The notable exception was the chimpanzee that showed a number of peaks in both multiplex assays, sometimes with high intensities and sometimes overlapping with the size ranges of the Y-STRs found in humans (Figure S7). Given the close genetic relationship between human and chimpanzee, with over 98% nucleotide identity [30], this is an expected finding. While RMplex 1 also displayed a small degree of cross-reactivity with the DNA samples from a cat, dog, rat, and a pig, RMplex 2 did not show any signs of cross-reactivity for any of these animal species. The electropherograms resulting from the non-human specificity testing can be found in Figure S7. Most comparable studies on different Y-STRs multiplex assays reported even less amplification of non-human DNA [13, 14, 31-34], hence there may be room for improvement on this aspect. However, contamination with non-human DNA in forensic casework would rather be the exception than the rule. Hence, the practical consequence of the somewhat lower human-specificity compared to other Y-STR kits is expected to be minor, at most.

Robustness

Two well-known PCR inhibitors, i.e., humic acid and tannic acid, were introduced to the PCR of both multiplexes at two different concentrations using varying amounts of input DNA. We saw that with higher amounts of input DNA, the assays were more resilient to the inhibition. However, the impact of these two inhibitors was quite significant, displaying incomplete profiles in the majority of these inhibition tests (Figure 5). The Y-

STRs that dropped out appear non-random. In RMplex 1, it was again DYS570 that dropped out most (71% dropout), which may be helped with the redesigned primer (see above), followed by DYS526b (42% dropout), DYR88 and DYS1013 (both 35% dropout), DYS713 (32% dropout), and DYF404S1 (23% dropout). The expected alleles for other Y-STRs, i.e., DYF393S1, DYS518, DYS627, DYS1003, and DYF1000 could be successfully detected in every experiment. Y-STRs that were especially sensitive to inhibition-based dropout in RMplex 2 were DYS449 (44% dropout), DYS547 (44% dropout), DYF403S1b (38% dropout), and DYS711 (16% dropout), whereas DYF399S1, DYS576, DYS612, showed no dropout in any of these experiments. As evident from Figure 5, the two PCR inhibitors tested showed slightly different impact on the two multiplex assays, such as humic acid having a larger negative impact on the performance of the assays compared to tannic acid.

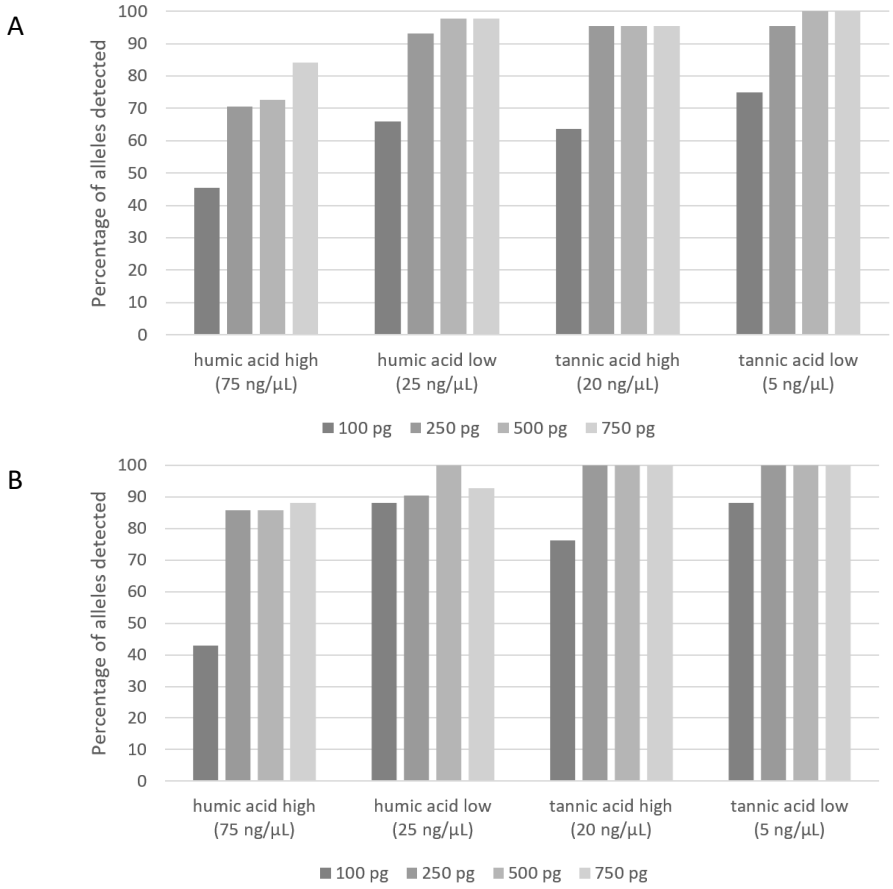


Figure 5: The effect of two well-known PCR inhibitors, humic acid and tannic acid, in different concentrations on different input DNA amounts with RMplex 1 (A) and RMplex2 (B).

Male-female mixtures

Given the typical forensic application of Y-STRs to male-female DNA mixtures from sexual assault cases, any Y-STR genotyping assay is required to be able to produce male-derived genotypes from male-female DNA mixtures with an excess amount of female DNA. A total of 14 mixtures were prepared at 7 different mixture ratios and with two different amounts of male input DNA; all mixtures were genotypes in triplicate with both genotyping assays. The results are summarized in Figure 6. It became evident that the performance of RMplex 2 in the presence of female DNA is highly superior to that of RMplex 1. In RMplex 1, allelic drop-out was observed starting in samples with a 1:20 mixture ratio. In contrast, full profiles could still be obtained in 1:200 mixtures while using RMplex 2, or even in 1:500 mixtures when using 250 pg of male input DNA. In RMplex 1, the number of alleles that dropped out increased rapidly with increasing amounts of female DNA present in the male-female mixture. The rather poor performance of RMplex 1 in these type of mixture samples may be explained by the non-male specific amplification products that were described in section 3.2.4. Hence, a solution could be, although not yet tested empirically, to exclude the primers targeting DYS1012 from the RMplex 1 reaction mix when dealing with highly unbalanced male-female mixtures, doing so could be expected to significantly increase the ability to detect alleles for the remaining Y-STRs.

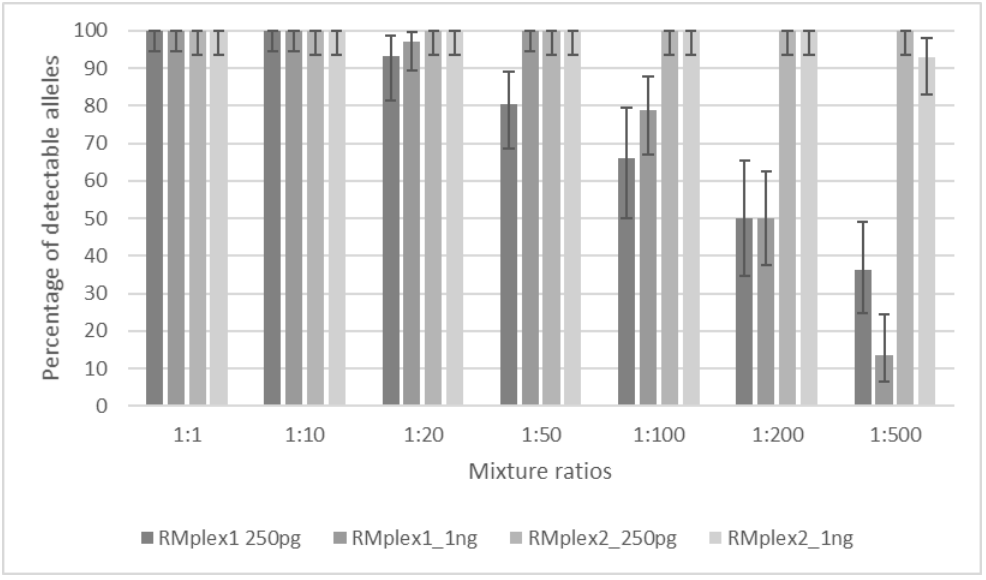


Figure 6: The performance of the two multiplex assays in male-female DNA mixtures; the error bars represent the exact binomial 95% confidence interval.

Degraded DNA

A total of eight samples with fragmented DNA were prepared. Their analysis with a 2100 Bioanalyzer estimated the average fragment length to be 684, 540, 477, 418, 337, 283, 205, and 183 bp, respectively. The samples with an average fragment size of 418 and larger resulted in complete Y-STR profiles, with a single exception of one replicate from the 477 bp sample that showed a single allele (DYF403S1b) slightly below the threshold value of 150 RFU. In fragmented samples with an average fragment length of 337 bp and lower, the performance of RMplex was increasingly affected by the shorter fragment lengths (Figure 7). There was a strong correlation with the size of the alleles and the number of allele that dropped out (R^2: 0.77, data not shown), where longer alleles dropped out more frequently than shorter ones. The performance of RMplex 1 appears to be slightly more affected compared to RMplex 2 when the input DNA is degraded (Figure 7).

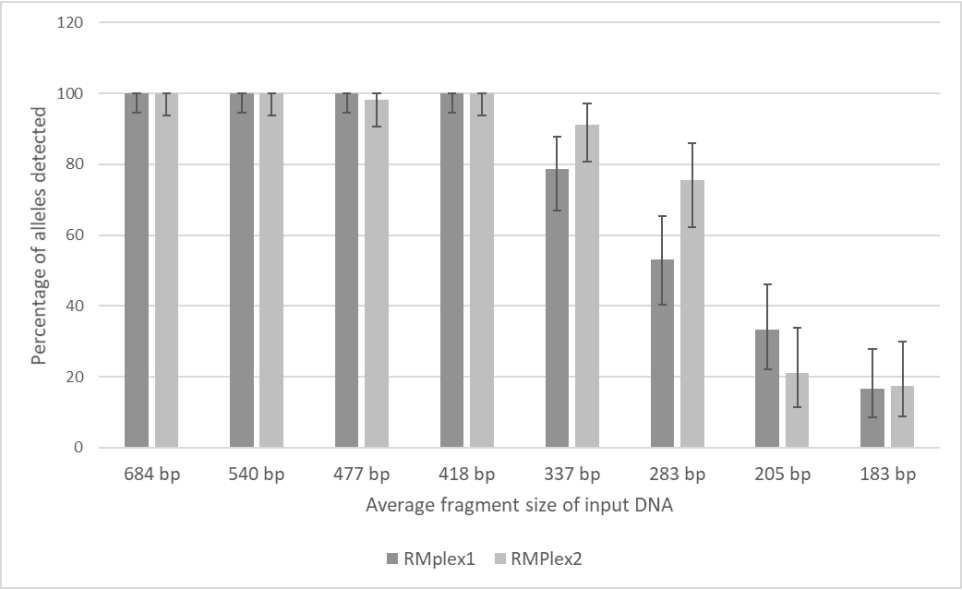


Figure 7: The performance of the two RMplex assays in samples with degraded DNA; the error bars represent the exact binomial 95% confidence interval.

Male relative differentiation

DNA samples from a total of 32 males belonging to five different paternal pedigrees in total were genotyped with both RMplex assays to obtain a first impression of the practical value of the set of 30 Y-STRs targeted by RMplex to differentiate male relatives. Surprisingly, when analyzing the genotypic data, it was noticed that there was a

remarkably large number of six out of 21 single-copy Y-STRs that exhibited two alleles in one of the individuals. Y-chromosomal duplications at so many loci in such a small sample set appears suspicious. Since the DNA of these individuals is derived from cell line material, we therefore regard it as more likely that these two alleles rather represent repeat mutation events during the cell propagation in cell culturing, which is supported by the increased mutation rates previously reported for these Y-STRs [1, 17]. Here, we took a conservative approach and interpreted such apparent duplicated alleles as no mutation, since we cannot be certain about the origin of these differences compared to the real mutation events we otherwise observed. For multi-copy Y-STRs we could not apply this conservative approach as germline mutations cannot easily be differentiated from mutations that occurred in culture, as we did when observing multiple alleles at a single-copy Y-STRs. Consequently, the number of genuine germline mutations in the multi-copy Y-STRs may be overestimated.

Following this conservative approach, a total of 23 repeat mutations were observed at 14 of the 30 Y-STRs analyzed. The largest number of mutations were observed at DYF1001 with five mutations, followed by DYF403S1a with three mutations and DYF399S1, DYF1000 and DYS518 all with two mutations each; notably, except DYS518, all of these are multi-copy Y-STRs. The Y-STRs with a single mutation observed were DYF393S1, DYS627, DYS713, DYS1003, DYF404S1, DYF387S1, DYS626, DYS612, and DYF1002. Twelve of the 23 mutations (52%) were found in one of the 13 initially described RM Y-STRs, another ten (43%) were seen on one of the 12 recently described RM Y-STRs, and a single mutation was found in one of the four included FM Y-STRs. These results may suggest that the set of 12 recently discovered RM Y-STRs indeed has a male relative differentiation capacity similar to that of the set of 13 previous RM Y-STRs, as was previously hypothesized [17]. However, the sample size here is at least an order of magnitude too small to make accurate claims. The same is true for the observed differentiation rates.

As shown in Figure S8, the 30 Y-STRs analyzed with the newly developed multiplex assays differentiated nearly 60% of the father-son pairs, over 80% of the brother pairs and over 75% of the grandfather-grandson pairs. Previous estimations on the differentiation capacity based on the mutation rates of the 26 RM Y-STRs included in the assays predicted that ~44% and father-son pairs and ~69% of brother pairs and grandfather-grandson pairs would be differentiated [17]. It has to be considered however, that the small sample size combined with stochastic effects and the effect of cell culturing may have led to an overestimation of the differentiation rates in the current study. Future,

more extensive studies, based on a large number of male relatives of different degrees, will have to confirm what the true empirical differentiation rate obtainable with RMplex is.

Lastly, we would like to emphasize that Y-STR haplotypes produced with this RMplex are expected to have extremely high discriminating capacity for unrelated males, on top of a relatively high discriminating capacity for related males. It is our opinion that extra caution is required when it comes to publishing haplotypic data based on such large number of Y-STRs with high mutation rates. In contrast to haplotypes based on standard Y-STRs with lower mutation rates generated using the previous Y-STR genotyping tools, Y-STR haplotypes established with RMplex can likely be matched to only a small number of related male individuals, and perhaps to single individuals. While potentially being the greatest value of this method within forensic casework, this same characteristic also poses potential privacy issues in a research setting. Therefore, we would like to recommend not to publish individual haplotypes that were produced using RMplex or other future methods with larger numbers of Y-STRs characterized by high mutation rates and only report about individual mutations that were observed, or overall population-based haplotype statistics.

Conclusion

Here we introduce RMplex: a novel multiplex genotyping method for the efficient analysis of 30 Y-STRs with high mutation rates based on both CE and MPS that covers all 26 currently known RM Y-STRs together with four FM Y-STRs and target a total of 44 male-specific loci. In the validation testing presented here, RMplex shows great promise for future applications to forensic practice. Provided successfully passing further forensic validation testing, we envision the application of RMplex in practical forensic casework. This will include criminal cases where is it suspected, or cannot be excluded, that instead of the suspect highlighted by matching the crime scene trace with standard Y-STRs, any of his male relatives may be the sample donor. RMplex could also be applied in Y-STR dragnets for familial search in serious criminal cases without known suspects, including cold cases, when matches with standard Y-STR kits are seen in too many volunteers. RMplex can then be used to separate the distant relatives, showing many mutations, from the close ones, with a few or no mutations, thereby allowing police investigation to focus on the close relatives, which provides increased chances to trace the unknown perpetrator. Lastly, RMplex is made available as efficient method to enlarge the empirical

data evidence for estimating mutation rates for and male relative differentiation rates based on these 30 Y-STRs by applying RMplex to large numbers of paternally related men of different degrees of known relationship in future research studies.

References

1. Ballantyne, K.N., et al., *Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications.* The American Journal of Human Genetics, 2010. **87**(3): p. 341-353.
2. Adnan, A., et al., *Improving empirical evidence on differentiating closely related men with RM Y-STRs: a comprehensive pedigree study from Pakistan.* Forensic Science International: Genetics, 2016. **25**: p. 45-51.
3. Boattini, A., et al., *Mutation rates and discriminating power for 13 rapidly-mutating Y-STRs between related and unrelated individuals.* PLOS One, 2016. **11**(11): p. e0165678.
4. Ballantyne, K.N., et al., *Toward Male Individualization with Rapidly Mutating Y-Chromosomal Short Tandem Repeats.* Human Mutation, 2014. **35**(8): p. 1021-1032.
5. Boattini, A., et al., *Estimating Y-Str Mutation Rates and Tmrca Through Deep-Rooting Italian Pedigrees.* Scientific Reports, 2019. **9**(1): p. 9032.
6. Zgonjanin, D., et al., *Mutation rate at 13 rapidly mutating Y-STR loci in the population of Serbia.* Forensic Science International: Genetics Supplement Series, 2017. **6**: p. e377-e379.
7. Čokić, V.P., et al., *A comprehensive mutation study in wide deep-rooted R1b Serbian pedigree: mutation rates and male relative differentiation capacity of 36 Y-STR markers.* Forensic Science International: Genetics, 2019. **41**: p. 137-144.
8. Serin, A., et al., *Genetic characterisation of 13 rapidly mutating Y-STR loci in 100 father and son pairs from South and East Turkey.* Annals of Human Biology, 2018. **45**(6-8): p. 506-515.
9. Javed, F., et al., *Male individualization using 12 rapidly mutating Y-STRs in Araein ethnic group and shared paternal lineage of Pakistani population.* International Journal of Legal Medicine, 2018. **132**(6): p. 1621-1624.
10. Zhang, W., et al., *Multiplex assay development and mutation rate analysis for 13 RM Y-STRs in Chinese Han population.* International Journal of Legal Medicine, 2017. **131**(2): p. 345-350.
11. Lang, M., et al., *Comprehensive mutation analysis of 53 Y-STR markers in father-son pairs.* Forensic Science International: Genetics Supplement Series, 2017. **6**: p. e57-e58.
12. Chen, Y., et al., *Mutation rates of 13 RM Y-STRs in a Han population from Shandong province, China.* Forensic Science International: Genetics Supplement Series, 2017. **6**: p. e346-e348.
13. Thompson, J.M., et al., *Developmental validation of the PowerPlex® Y23 System: a single multiplex Y-STR analysis system for casework and database samples.* Forensic Science International: Genetics, 2013. **7**(2): p. 240-250.
14. Gopinath, S., et al., *Developmental validation of the Yfiler® Plus PCR Amplification Kit: An enhanced Y-STR multiplex for casework and database applications.* Forensic Science International: Genetics, 2016. **24**: p. 164-175.
15. Alghafri, R., et al., *A novel multiplex assay for simultaneously analysing 13 rapidly mutating Y-STRs.* Forensic Science International: Genetics, 2015. **17**: p. 91-98.
16. Lee, E.Y., et al., *A multiplex PCR system for 13 RM Y-STRs with separate amplification of two different repeat motif structures in DYF403S1a.* Forensic Science International: Genetics, 2017. **26**: p. 85-90.

17. Ralf, A., et al., *Identification and characterization of novel rapidly mutating Y-chromosomal short tandem repeat markers.* Human Mutation, 2020. **41**(9): p. 1680-1696.
18. Cunningham, F., et al., *Ensembl 2019.* Nucleic acids research, 2019. **47**(D1): p. D745-D751.
19. Untergasser, A., et al., *Primer3—new capabilities and interfaces.* Nucleic acids research, 2012. **40**(15): p. e115-e115.
20. Tusnady, G.E., et al., *BiSearch: primer-design and search tool for PCR on bisulfite-treated genomes.* Nucleic Acids Research, 2005. **33**(1): p. e9-e9.
21. Shen, Z., et al., *MPprimer: a program for reliable multiplex PCR primer design.* BMC bioinformatics, 2010. **11**(1): p. 143.
22. Zhang, J., et al., *PEAR: a fast and accurate Illumina Paired-End reAd mergeR.* Bioinformatics, 2014. **30**(5): p. 614-620.
23. Gaspar, J.M., *NGmerge: merging paired-end reads via novel empirically-derived models of sequencing errors.* BMC bioinformatics, 2018. **19**(1): p. 1-9.
24. Hoogenboom, J., T. Sijen, and K.J. van der Gaag, *STRNaming: Generating simple, informative names for sequenced STR alleles in a standardised and automated manner.* Forensic Science International: Genetics, 2021: p. 102473.
25. Zhang, G.Q., et al., *Structure and polymorphism of 16 novel Y-STRs in Chinese Han Population.* Genetics and Molecular Research, 2012. **11**(4): p. 4487-4500.
26. Brookes, C., et al., *Characterising stutter in forensic STR multiplexes.* Forensic Science International: Genetics, 2012. **6**(1): p. 58-63.
27. Ravasini, F., et al., *Sequence read depth analysis of a monophyletic cluster of Y chromosomes characterized by structural rearrangements in the AZFc region resulting in DYS448 deletion and DYF387S1 duplication.* Frontiers in Genetics, 2021. **12**.
28. Watahiki, H., et al., *Differences in DYF387S1 copy number distribution among haplogroups caused by haplogroup-specific ancestral Y-chromosome mutations.* Forensic Science International: Genetics, 2020. **48**: p. 102315.
29. Gusmão, L., et al., *DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis.* Forensic science international, 2006. **157**(2-3): p. 187-197.
30. Hughes, J.F., et al., *Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content.* Nature, 2010. **463**(7280): p. 536-539.
31. Mo, X.-T., et al., *Developmental validation of the DNATyper™ Y26 PCR amplification kit: an enhanced Y-STR multiplex for familial searching.* Forensic Science International: Genetics, 2019. **38**: p. 113-120.
32. Du, W., et al., *developmental validation of a novel 6-dye typing system with 36 Y-STR loci.* International journal of legal medicine, 2019. **133**(4): p. 1015-1027.
33. Li, M., et al., *Development and validation of a novel 29-plex Y-STR typing system for forensic application.* Forensic Science International: Genetics, 2020. **44**: p. 102169.
34. Bai, R., et al., *Developmental Validation of a novel 5 dye Y-STR System comprising the 27 YfilerPlus loci.* Scientific reports, 2016. **6**(1): p. 1-8.

Supplementary information

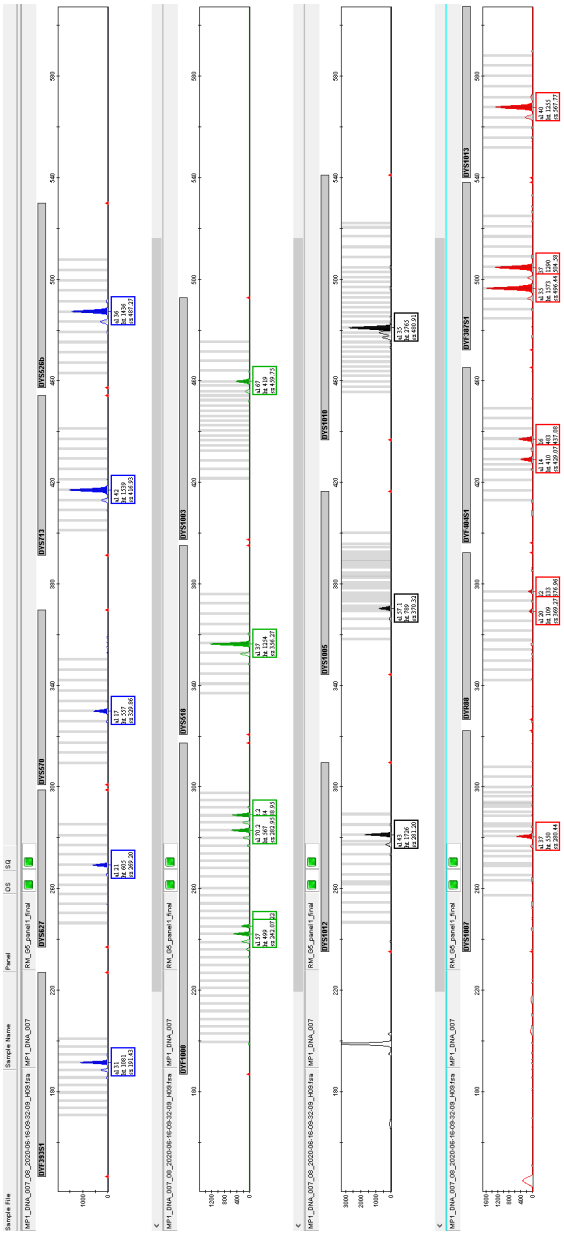


Figure S1: Electropherogram of AmpFℓSTR DNA Control 007 analyzed with multiplex 1.

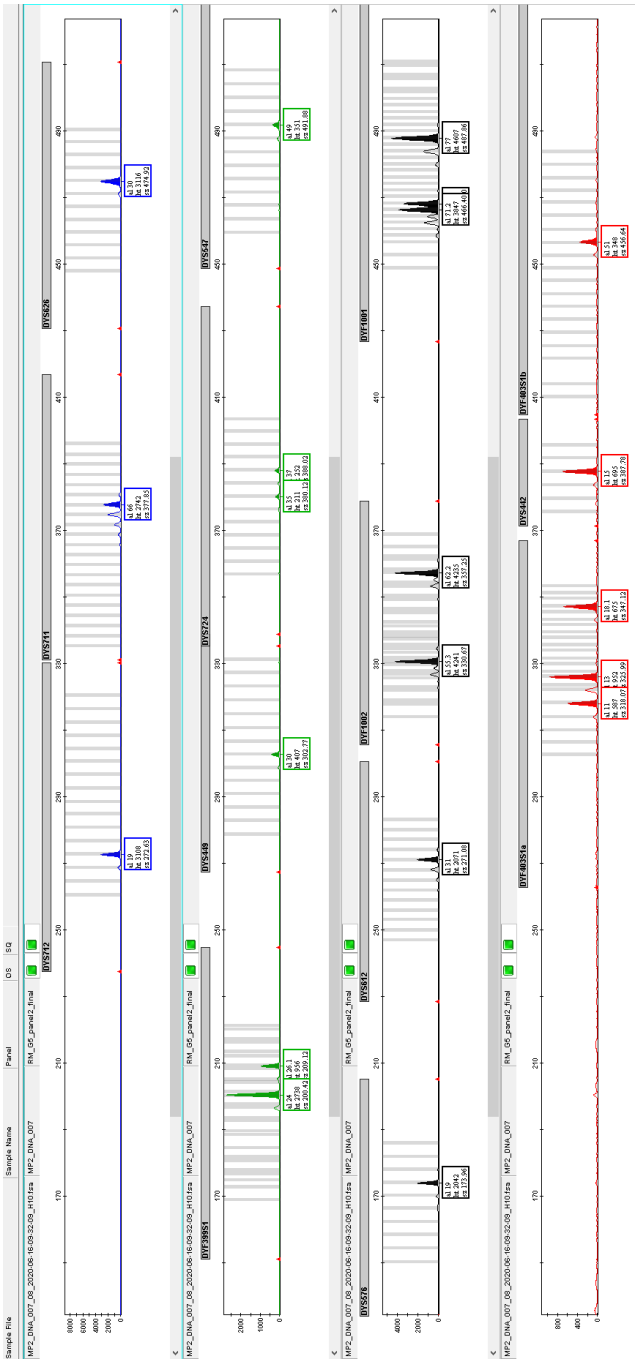


Figure S2: Electropherogram of AmpFℓSTR DNA Control 007 analyzed with multiplex 2.

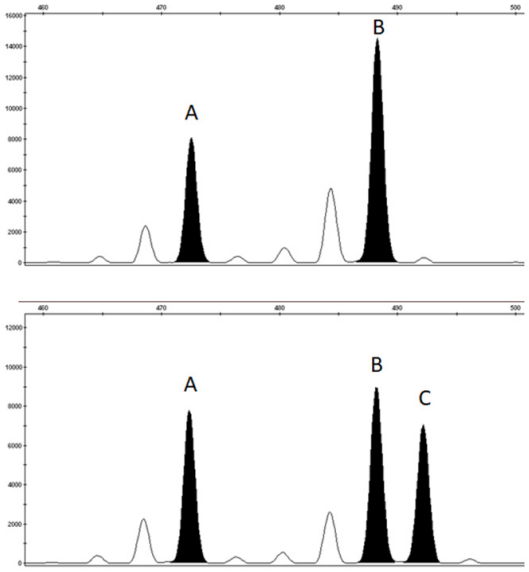


Figure S3: An example of a multi-copy Y-STR DYF1001 tested in two paternally related males, showing a B > C or C > B allele mutation depending on their relationship .

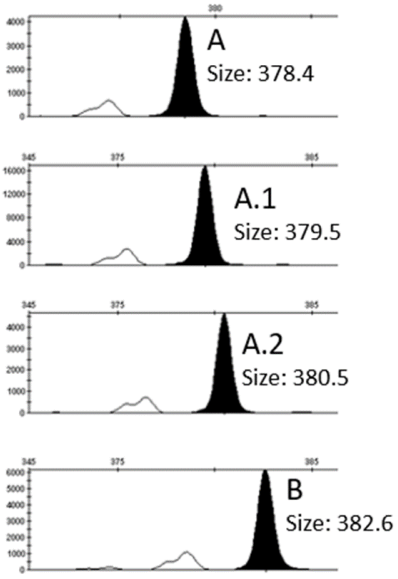


Figure S4: Different microvariant alleles observed at the predominantly tetranucleotide Y-STR DYS1005 as likely caused by mutations in the pentanucleotide repeat stretch this Y-STR additionally contains.

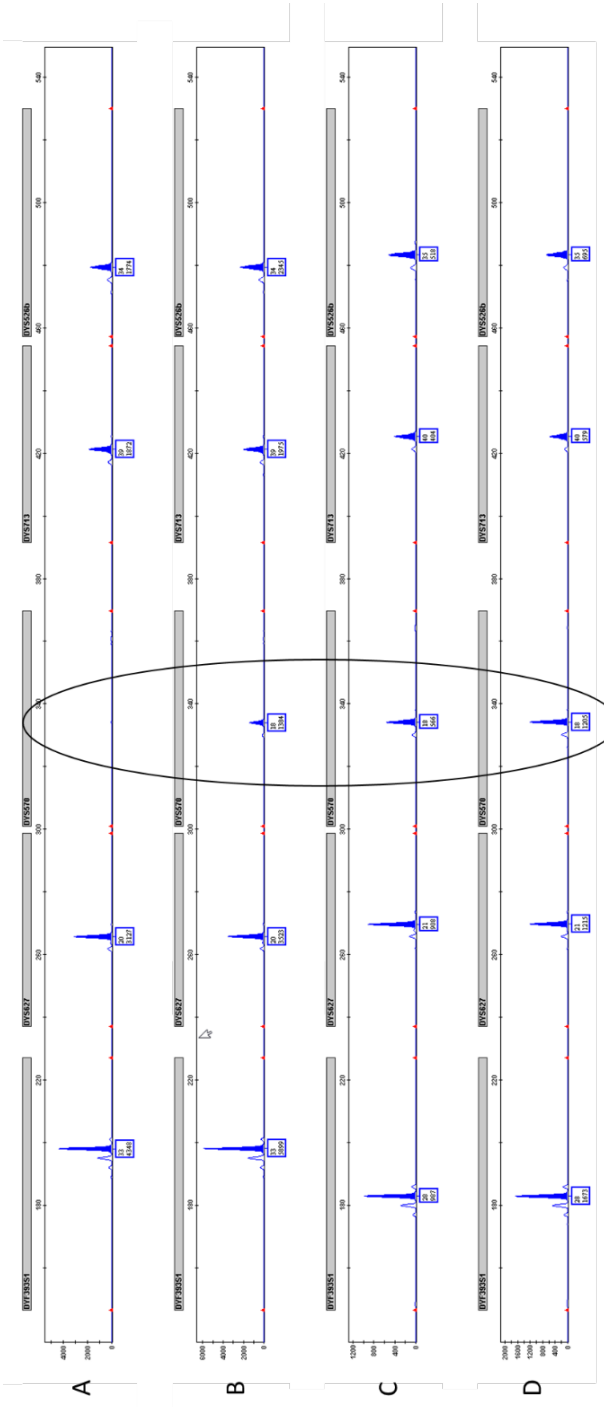


Figure S5: Electropherogram of the Fam-channel of RMplex 1 in a haplogroup E male with the original primer (A) and the modified primer (B) for DYS570 and a non-haplogroup E male with the original (C) and the modified (D) primer for DYS570.

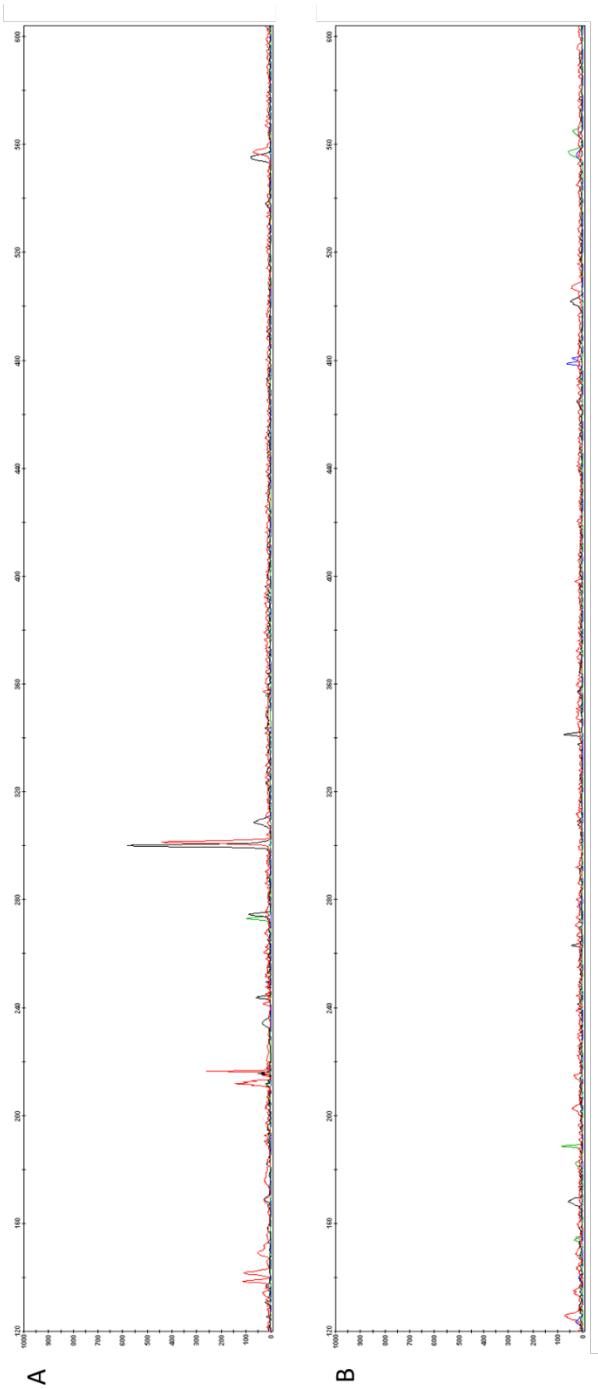


Figure S6: Electropherograms of a female sample after amplification with RMplex 1 (A) and RMplex 2 (B)

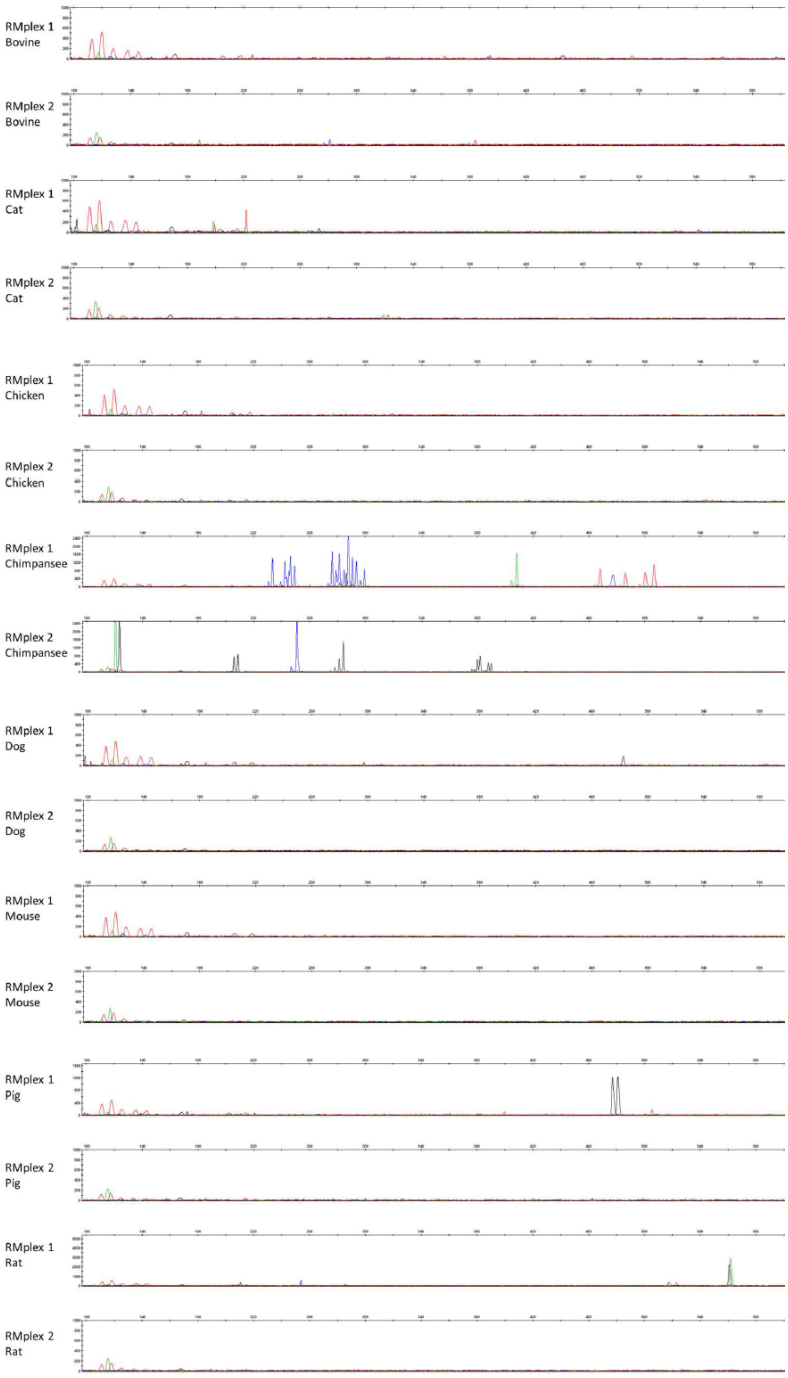


Figure S7: Electropherograms of animal samples after amplification with RMplex 1 and RMplex 2.

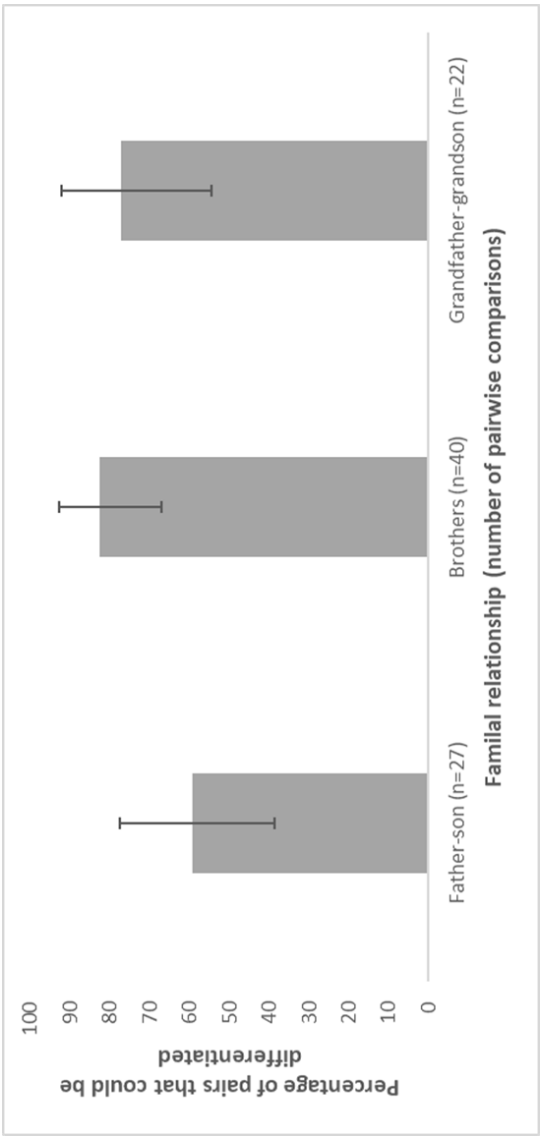


Figure S8: Preliminary male relative differentiation rates based on 30 Y-STRs targeted with RMplex from five CEPH pedigrees, the error bars represent the binomial 95% confidence intervals.

Supplementary Table S1: Genomic locations of all Y-STRs

Y-STR	GRCh38 location
DYF393S1	chrY:21599660-21599853
DYS627	chrY:8781895-8782172
DYS570	chrY:6993020-6993352
DYS713	chrY:7963883-7964301
DYS526b	chrY:3772235-3772728
DYF1000	chrY:17854636-17854920; chrY:18051087-18051374; chrY:24024241-24024481; chrY:25645652-25645895
DYS518	chrY:15207925-15208293
DYS1003	chrY:15084578-15085044
DYS1012	chrY:56858298-56858581
DYS1005	chrY:7849473-7849841
DYS1010	chrY:17324692-17325179
DYS1007	chrY:17299965-17300246
DYR88	chrY:24331940-24332308; chrY:25337841-25338209
DYF404S1	chrY:23807705-23808142; chrY:25861946-25862375
DYF387S1	chrY:23785179-23785675; chrY:25884405-25884905
DYS1013	chrY:14697088-14697659
DYS712	chrY:13446511-13446799
DYS711	chrY:8465382-8465755
DYS626	chrY:22270727-22271211
DYF399S1	chrY:22950235-22950447; chrY:24583992-24584191; chrY:25085898-25086101
DYS449	chrY:8349870-8350170
DYS724	chrY:24005762-24006153; chrY:25663983-25664378
DYS547	chrY:16759991-16760484
DYS576	chrY:7185278-7185452
DYS612	chrY:13640641-13640912
DYF1002	chrY:16242212-16242558; chrY:16342765-16343126
DYF1001	chrY:23054574-23055066; chrY:24688314-24688812; chrY:24981295-24981785
DYF403S1a	chrY:6357792-6358116; chrY:9681990-9682343; chrY:9816611-9816939
DYS442	chrY:12649041-12649429
DYF403S1b	chrY:6479576-6480025

Supplementary Table S2: Assigning the repeat numbers to GRCh38

Marker name	Sequence length	STRnaming nomenclature[#]	Rationale	Inferred repeat number	Final repeat number	Source final repeat numbers
DYF393S1	193	CE27_GAA**[4]**AAG**[23]**CAG**[1]**AAG**[1]**	4+23+1+1	29	29	This study
DYS627	277	CE19_GA[3]CACA[1]GGGA[1]GAAA**[3]**GGAAAA[1]AAGA**[3]**GA[4]GAAA**[19]**AAGA[1]GAAA[1]GA[3]GGGA[2]	3+3+19	25	21	YF Plus
DYS570	332	CE17_**TTTC[17]**	17	17	17	PPY23+YF Plus
DYS713	418	CE20_**TTTC[18]**TCTT[2]TCTG[1]TCTT[1]T[1]TTTC[2]**TCTT[12]TTTC[5]**[]TTAT[1]**TATT[4]**	18+12+5+4	39	42	Zhang et al. 2012
DYS526b	493	CE40_CTTC[1]**CCTC[3]**TTTC[1]TTCC[1]TTTC[2]**TTCC[1]TTTC[13]CTTC[9]**TTTC[1]CTTC[1]TTTC[1]CTTC[1]ATTT[1]CTTC[1]TTTC[2][]**TCCT[14]**	3+1+13+9+14	40	37	Ballantyne et al. 2014
DYF1000_1	240	CE57_AAG**[9]**CAG[1]AAG**[22]**AGG**[1]**AAG**[1]**AGG**[1]**AAG**[22]**	9+22+1+1+1+22	56	56	This study
DYF1000_2*	243	CE58_AAG**[9]**CAG[1]AAG**[24]**AGG**[1]**AAG**[1]**AGG**[1]**AAG**[21]**	9+24+1+1+1+21	57	57	This study
DYF1000_3	284	CE33_AAG**[5]**AAGGAG**[6]**AAG[3]GAG[3]GA[1]GAA**[5]**GGA[2]AGA**[5]**AGG[1]AGA[1]GAA[2]G[1]GAA**[4]**GAT[1]GAA[2]AGA**[26]**	5+12+5+5+4+26	57	70.2	This study
DYF1000_4*	287	CE33_AAG**[5]**AAGGAG**[6]**AAG[3]GAG[3]GA[1]GAA**[6]**GGA[2]AGA**[5]**AGG[1]AGA[1]GAA[2]G[1]GAA**[4]**GAT[1]GAA[2]AGA**[26]**	5+12+6+5+4+26	58	71.2	This study
DYS518	368	CE24_GAAA[2]GGAA[1]GAAA**[16]**GGGA[1]GAAA**[4]**GGAA[1]GA[1]GAAA**[13]**GGAA[1]GA[1]GGGA[2]AGAA[1]GAAA[1]AGAA[2]**GGAA[3]**	16+4+13+3	36	38	YF Plus
		CE42_AAAG[1]AAGA[2]AGGA[1]AAGA**[3]**AAGG[1]AACG[1]AAGG**[11]**AAGA**[3]**GAAA[3]AAAG[1]AGGG**[3]**AAGG[1]AGGA**[3]**[]	3+11+3+3+3+3+18+			
DYS1003	466	GA[4]ACGG[1]AAGA**[18]**AAGG[1]TAGA[1]AAGA**[14]**GAAA**[5]**AAGA**[3]**	14+5+3	66	66	This study
		CE34_AAGA**[14]**AAGG[1]AAAG[1]AA[1]AAGA[2]AAAG**[3]**AA[1]AAGA**[5]**AAAG**[3]**AAGG[1]AAAG[1]AAGG[1]AAAG[1]AAGG[1]AAAG[1]				
DYS1012	283	AAGG[1]AAAG[1]AAGG[2]AAAG**[16]**	14+3+5+3+16	41	41	This study
DYS1005	368	CE12_**TTCCT[12]TCCC[7]**TTCT[2]**CTTT[5]CCTT[13]CTTT[12]**CTT[1]**CCTT[4]**CTTT[2]	12+7+5+13+12+4	53	53	This study
DYS1010	487	CE17_AAGA**[17]**GA**[13]**AAGA**[11]**AAA[1]AAGA[1]GAGG[2]GA[3]	17+6.2+11	34.2	34.2	This study
DYS1007	281	CE37_AAGA**[14]**GAAA**[3]**GAGA[1]AAGA**[5]**GAAA**[14]**[]GAAA[1]GA[5]CAGC[1]AAGA[1]GAAA[1]GA[5]	14+3+5+14	36	36	This study
DYR88_1	368	CE19_AAAG**[19]**	19	19	19	This study
DYR88_2*	368	CE19_AAAG**[19]**	19	19	19	This study
DYF404S1_1*	429	CE14_**TTTC[14]**TCTC[1]TTTC[1]	14	14	14	Ballantyne et al. 2014
DYF404S1_2	437	CE16_**TTTC[16]**TCTC[1]TTTC[1]	16	16	16	Ballantyne et al. 2014
DYF387S1_1	496	CE35_A[7]GAA**[4]**AGAG[1]A[7]GAAA[2]GGTA[1]GGAA**[4]**GAAA[2]GGAA[1]GAAA[2]GGAA**[9]**GAAA**[14]**AT[1]A[8]	3+4+9+14	30	35	YF Plus
DYF387S1_2*	500	CE36_A[7]GAA**[4]**AGAG[1]A[7]GAAA[2]GGTA[1]GGAA**[4]**GAAA[2]GGAA[1]GAAA[2]GGAA**[10]**GAAA**[14]**AT[1]A[8]	3+4+10+14	31	36	YF Plus
DYS1013	571	CE13_AGAT**[11]**TGAT[1]AGAT[1][]AGAT[2]AGAA[1]AGAT**[6]**[]GATA**[11]**[]AGAT**[10]**AGGC**[3]**	11+6+11+10+3	41	41	This study
DYS712	288	CE22_AGAT**[17]**AGAC**[5]**	17+5	22	22	Zhang et al. 2012
DYS711	373	CE34_**TTC[4]**T[1]**TTC[31]**TCC[3]TTC[1]**[]CTT[25]**CTCCTC[1]CTT[3]	4+31+25	60	64	Zhang et al. 2012
DYS626	484	CE17_AAAG[1]A[9]GCAG[1]AAAG[1]GAAG[1]A[5]GAAA**[17]**[]GAAG[3]AAAG**[3]**AGACAG[1]AAAG**[4]**A[6]GAAA[2]GAAG[1]GAAA**[3]**	17+3+3+4+3	30	30	Ballantyne et al. 2014
DYF399S1_1	199	CE18_AAAG**[3]**AAAAAGA[1]AAAG**[18]**	3+18	21	22	Ballantyne et al. 2014
DYF399S1_2*	203	CE19_AAAG**[3]**AAAAAGA[1]AAAG**[18]**_+2.1->AAAA_+19A>G	3+18+1	22	23	Ballantyne et al. 2014
DYF399S1_3	212	CE21.1_AAAG**[3]**AAAAAGA[1]AAAGG**[1]**AAAG**[19]**AAAA**[1]**	3+1.1+19+1	24.1	25.1	Ballantyne et al. 2014
DYS449	300	CE23_**TCTT[14]**TC[4]TCC[2]TCTT[2]TCCT[1]**TCTT[3]**TTCC[1]TCTT[1]TCCT[1]**TCTT[14]**	14+3+14	31	29	YF Plus
DYS724_1	391	CE20_CTTT[2]**TTTC[4]**TCTC[1]TTTC[2]TC[5]**CCTT[17]**T[1]CCTT[1]**CTTT[19]**	4+17+19	40	36	FT-DNA
DYS724_2*	395	CE21_CTTT[2]**TTTC[4]**TCTC[1]TTTC[2]TC[5]**CCTT[16]**T[1]CCTT[1]**CTTT[21]**	4+16+21	41	37	FT-DNA
		CE24_**CTTC[3]**TTTC[1]**CTTC[11]**CTTT[1]**CTTC[5]**TTTC[1]CTCC[2]**[]**CTCC[1]CTTC[1]CTCC[1]TTCC[1]CTCC[2]TTTC[1]T[5]**CTTT[15]**CCTTA[1]				
DYS547	493	TTTC[1]**TTCC[4]**TTTC[1]**TCTT[12]**TTCT[2]CTCT[1]**TTTC[3]**T[4]CTTT[2]T[6]	3+11+5+15+4+12+3	53	49	Ballantyne et al. 2014
DYS576	174	CE17_AAAG**[17]**	17	17	17	PPY23+YF Plus
DYS612	271	CE36_CCT**[5]**CTT[1]**TCT[4]**CCT[1]**TCT[25]**	5+4+25	34	30	Forenseq
		CE35_GA**[7]**GAAA**[4]**TAGA[1]AAGA**[13]**GAAA**[4]**GGAA[1]AGAA**[12]A[1]**CAGAGA[1]AAGA**[4]**GAAA**[10]**AAGA[1]A[1]AAGA[2]	3.2+4+13+4+12+0.1+			
DYF1002_1	346	GGAA[1]GAAA**[4]**AAGA[1]GAAA[2]GGAG**[3]**GGAA[2]AGAA[1]GGAA[1]GGAG[1]	4+10+4+3	57.3	57.3	This study
		CE38.3_GA**[7]**GAAA**[4]**TAGA[1]AAGA**[12]**AAGG**[1]**GAAA**[5]**GGAA[1]AGAA**[10]**GAAA**[2]**CAGAGA[1]AAGA**[4]**GAAA**[13]**	3.2+4+12+1+5+10+2+4+			
DYF1002_2*	361	AAGA[1]A[1]AAGA[2]GGAA[1]GAAA**[4]**AAGA[1]GAAA[2]GGAG**[3]**GGAA[2]AGAA[1]GGAA[1]GGAG[1]	13+4+3	61.2	61.2	This study
		CE23_GAAG**[10]**[]AAGA**[15]**GA**[2]**AAGA**[4]**GAAA[1]**GA[4]**GAAA[1]GA[3]AAATGAAAGAAAAAGAA[1]GGAA**[6]**GAAA**[14]**TAAA**[1]**	10+15+1+4+2+6+14+1+			
DYS1001_1*	490	AGAA[1]AAGA**[4]**AGGA**[3]**AAGA[1]GAAA**[3]**GGAA**[1]**GAAA**[1]**GGAA**[10]**	4+3+3+1+1+10	75	75	This study
		CE23.2_GAAG**[9]**[]AAGA**[15]**GAAA**[1]**GA**[3]**AAGA**[4]**GAAA[1]GA**[4]**GAAA[1]GA[3]AAATGAAAGAAAAAGAA[1]GGAA**[4]**GAAA**[14]**	9+15+1+1.2+4+2+4+14+			
DYS1001_2	492	AGAA[1]AAGA**[4]**AGGA**[3]**AAGA[1]GAAA**[3]**GGAA**[15]**	4+3+3+15	75.2	75.2	This study
		CE25_GAAG**[12]**[]AAGA**[16]**GA**[2]**AAGA**[4]**GAAA[1]GA**[4]**GAAA[1]GA[3]AAATGAAAGAAAAAGAA[1]GGAA**[6]**GAAA**[12]**TAAA**[1]**	12+16+1+4+2+6+12+1+			
DYS1001_3	498	AGAA[1]AAGA**[4]**AGGA**[3]**AAGA[1]GAAA**[3]**GGAA**[1]**GAAA**[1]**GGAA**[11]**	4+3+3+1+1+11	77	77	This study
DYF403S1a_1	324	CE15_TTCT[2]TTCC[1]**TTCT[13]**ATC[1]**TTCT[13]**	13+13	26	11	Ballantyne et al. 2014
DYF403S1a_2*	328	CE16_TTCT[2]TTCC[1]**TTCT[10]**ATC[1]**TTCT[17]**	10+17	27	12	Ballantyne et al. 2014
DYF403S1a_3*	353	CE28_TTTC[2]GTTC[1]TCTT[2]TTTC[1]TTTT[1]**TCTT[17]TTTC[4]**	17+4	21	18.1	Ballantyne et al. 2014
DYS442	388	CE17_GATA**[12]**GACA**[3]**GATA[2]	12+3	15	15	This study
DYF403S1b	449	CE53_GAAA**[4]**AAGA**[14]**AGGA**[9]**AAGA**[9]**GAAA**[12]**AAGA[2]GAAC[1]GAAA[2]	4+14+9+9+12	48	47	Ballantyne et al. 2014

* Used reverse complement sequence for compatibility

[#] Sequence blocks displayed in bold are used for the inferred repeat number

Supplementary Table S3: Assigning the repeat numbers to AmpF€STR DNA Control 007

Marker name	CE Length	MPS Length	STRnaming nomenclature[#]	Rationale	Inferred repeat number	Final repeat number	Source final repeat numbers
DYF393S1	191.4	199	CE29_GAA**[4]**AAG**[25]**CAG**[1]**AAG**[1]**	4+25+1+1	31	31	This study
DYS627	269.2	277	CE19_GA[3]CACA[1]GGGA[1]GAAA**[3]**GGAAAA[1]AAGA**[3]**GA[4]GAAA**[19]**AAGA[1]GAAA[1]GA[3]GGGA[2]	3+3+19	25	21	YF Plus
DYS570	329.9	332	CE17_**TTTC[17]**	17	17	17	PPY23+YF Plus
DYS713	416.9	418	CE20_**TTTC[18]**TCTT[2]TCTG[1]TCTT[1]T[1]TTTC[2]**TCTT[12]TTTC[5]**[]TTAT[1]**TATT[4]**	18+12+5+4	39	42	Zhang et al. 2012
DYS526b	487.3	489	CE39_CTTC[1]**CCTC[3]**TTTC[1]TTCC[1]TTTC[2]**TTCC[1]TTTC[13]CTTC[9]**TTTC[1]CTTC[1]TTTC[1]CTTC[1]ATTT[1]CTTC[1]TTTC[2][]**TCCT[13]**	3+1+13+9+13	39	36	Ballantyne et al. 2014
DYF1000_1	242.1	243	CE58_AAG**[9]**CAG[1]**AAG[22]AGG[1]AAG[1]AGG[1]AAG[23]**	9+22+1+1+1+23	57	57	This study
DYF1000_2	245.2	246	CE59_AAG**[9]**CAG[1]**AAG[22]AGG[1]AAG[1]AGG[1]AAG[24]**	9+22+1+1+1+24	58	58	This study
DYF1000_3	282.8	284	CE33_AAG**[5]**AAGGAG**[6]**AAG[3]GAG[3]GA[1]GAA**[6]**GGA[2]AGA**[5]**AGG[1]AGA[1]GAA[2]G[1]GAA**[4]**GAT[1]GAA[2]AGA**[25]**	5+12+6+5+4+25	57	70.2	This study
DYF1000_4	289.0	290	CE34_AAG**[5]**AAGGAG**[6]**AAG[3]GAG[3]GA[1]GAA**[5]**GGA[2]AGA**[5]**AGG[1]AGA[1]GAA[2]G[1]GAA**[4]**GAT[1]GAA[2]AGA**[28]**	5+12+5+5+4+28	59	72.2	This study
DYS518	356.3	364	CE23_GAAA[2]GGAA[1]GAAA**[15]**GGGA[1]GAAA**[4]**GGAA[1]GA[1]GAAA**[13]**GGAA[1]GA[1]GGGA[2]AGAA[1]GAAA[1]AGAA[2]GGAA**[3]** CE43_AAAG[1]AAGA[2]AGGA[1]**AAGA[3]**AAGG[1]AACG[1]**AAGG[11]AAGA[3]GAAA[3]**AAAG[1]AGGG[3]AAGG[1]AGGA[3][]GA[4]	15+4+13+3 3+11+3+3+3+3+	35	37	YF Plus
DYS1003	459.8	470	ACGG[1]**AAGA[19]**AAGG[1]TAGA[1]AAGA**[14]**GAAA**[5]AAGA[3]** CE36_**AAGA[15]**AAGG[1]AAAG[1]AA[1]AAGA[2]AAA**G[3]**AA[1]**AAGA[5]AAAG[3]**AAGG[1]AAAG[1]AAGG[1]AAAG[1]AAGG[1]AAAG[1]	19+14+5+3	67	67	This study
DYS1012	281.2	291	AAGG[1]AAAG[1]AAGG[2]**AAAG[17]**	15+3+5+3+17	43	43	This study
DYS1005	370.3	373	CE13_**TTCCT[13]TCCC[7]**TTCT[2]**CTTT[5]CCTT[14]CTTT[11]**CTT[1]C**CTT[4]**CTTT[2]	16.1+7+5+14+11+4	57.1	57.1	This study
DYS1010	480.9	489	CE17.2_**AAGA[17]GA[14]AAGA[11]**AAA[1]AAGA[1]GAGG[2]GA[3]	17+7+11	35	35	This study
DYS1007	280.4	285	CE38_**AAGA[16]GAAA[2]**GAGA[1]**AAGA[5]GAAA[14]**[]GAAA[1]GA[5]CAGC[1]AAGA[1]GAAA[1]GA[5]	16+2+5+14	37	37	This study
DYR88_1	369.3	372	CE20_**AAAG[20]**	20	20	20	This study
DYR88_2	377.0	380	CE22_**AAAG[22]**	22	22	22	This study
DYF404S1_1	429.1	429	CE14_**TTTC[14]**TCTC[1]TTTC[1]	14	14	14	Ballantyne et al. 2014
DYF404S1_2	437.1	437	CE16_**TTTC[16]**TCTC[1]TTTC[1]	16	16	16	Ballantyne et al. 2014
DYF387S1_1	496.4	496	CE35_A[7]**GAA[4]**AGAG[1]A[7]GAAA[2]GGTA[1]GGAA[4]GAAA[2]GGAA[1]GAAA[2]**GGAA[9]**GAAA**[14]**AT[1]A[8]	3+4+9+14	30	35	YF Plus
DYF387S1_2	504.6	504	CE37_A[7]**GAA[4]**AGAG[1]A[7]GAAA[2]GGTA[1]GGAA**[4]**GAAA[2]GGAA[1]GAAA[2]**GGAA[10]GAAA[15]**AT[1]A[8]	3+4+10+15	32	37	YF Plus
DYS1013	567.8	567	CE12_**AGAT[11]**TGAT[1]AGAT[1][]AGAT[2]AGAA[1]**AGAT[6]**[]**GATA[10]**[]**AGAT[10]AGGC[3]**	11+6+10+10+3	40	40	This study
DYS712	272.6	276	CE19_**AGAT[14]AGAC[5]**	14+5	19	19	Zhang et al. 2012
DYS711	377.7	379	CE36_**TTC[4]**T[1]**TTC[33]**TCC[3]TTC[1]**[]CTT[25]**CTC[2]CTT[3]	4+33+25	62	66	Zhang et al. 2012
DYS626	474.9	484	CE17_AAAG[1]A[9]GCAG[1]AAAG[1]GAAG[1]A[5]GAAA**[17]**[]GAAG**[3]**AAAG**[3]**AGACAG[1]AAAG**[4]**A[6]GAAA[2]GAAG[1]GAAA**[3]**	17+3+3+4+3	30	30	Ballantyne et al. 2014
DYF399S1_1	200.4	207	CE20_**AAAG[3]**AAAAAGA[1]AAAG**[20]**	3+20	23	24	Ballantyne et al. 2014
DYF399S1_2	200.4	207	CE20_**AAAG[3]**AAAAAGA[1]AAAG**[19]**_+2.1->AAAA_+19A>G	3+19+1	23	24	Ballantyne et al. 2014
DYF399S1_3	209.1	216	CE22.1_AAAG**[3]**AAAAAGA[1]AAAGG**[1]AAAG[20]AAAA[1]**	3+1.1+20+1	25.1	26.1	Ballantyne et al. 2014
DYS449	302.8	304	CE24_**TCTT[15]**TC[4]TCC[2]TCTT[2]TCCT[1]**TCTT[3]**TTCC[1]TCTT[1]TCCT[1]**TCTT[14]**	15+3+14	32	30	YF Plus
DYS724_1	380.1	387	CE19_CTTT[2]**TTTC[4]**TCTC[1]TTTC[2]TC[5]**CCTT[16]**T[1]CCTT[1]**CTTT[19]**	4+16+19	39	35	FT-DNA
DYS724_2	388.0	395	CE21_CTTT[2]**TTTC[4]**TCTC[1]TTTC[2]TC[5]**CCTT[17]**T[1]CCTT[1]**CTTT[20]** CE24_**CTTC[3]TCCT[1]**TTTC[1]**CTTC[11]**CTTT[1]**CTTC[5]**TTTC[1]CTCC[2][]CTCC[1]CTTC[1]CTCC[1]TTCC[1]CTCC[2]TTTC[1]T[5]**CTTT[15]**	4+17+20	41	37	FT-DNA
DYS547	491.9	493	CCTTA[1]TTTC[1]**TTCC[4]**TTTC[1]**TCTT[11]**TTCT[2]CTCT[1]**TTTC[3]**T[4]CTTT[2]T[6]	3+1+11+5+15+4+11+3	53	49	Ballantyne et al. 2014
DYS576	174.0	182	CE19_**AAAG[19]**	19	19	19	PPY23+YF Plus
DYS612	271.1	274	CE37_CCT**[5]**CTT[1]**TCT[4]**CCT[1]**TCT[26]** CE33_GA**[7]GAAA[4]**TAGA[1]**AAGA[11]**GAAA[4]GGAA[1]AGAA**[12]A[1]**CAGAGA[1]**AAGA[4]GAAA[10]**AAGA[1]A[1]AAGA[2]GGAA[1]	5+4+26 3.2+4+11+4+12+0.1+	35	31	Forenseq
DYF1002_1	330.7	338	GAAA**[4]**AAGA[1]GAAA[2]GGAG**[3]**GGAA[2]AGAA[1]GGAA[1]GGAG[1] CE39.3_GA**[7]**GAAA**[4]**TAGA[1]**AAGA[12]AAGG[1]GAAA[5]**GGAA[1]AGAA**[10]**GAAA[2]CAGAGA[1]**AAGA[4]GAAA[14]**AAGA[1]A[1]	4+10+4+3 3.2+4+12+1+5+10+2+	55.3	55.3	This study
DYF1002_2	357.3	365	AAGA[2]GGAA[1]GAAA**[4]**AAGA[1]GAAA[2]GGAG**[3]**GGAA[2]AGAA[1]GGAA[1]GGAG[1] CE19.2_GAAG**[9]**[]**AAGA[13]GAAA[1]GA[3]AAGA[4]**GAAA[1]GA**[4]**GAAA[1]GA[3]AAATGAAAGAAAAAGAA[1]GGAA**[4]GAAA[12]**	4+14+4+3 9+13+1+1.2+4+2+4+	62.2	62.2	This study
DYS1001_1	466.4	476	AGAA[1]AAGA**[4]**AGGA**[3]**AAGA[1]GAAA**[3]GGAA[15]** CE20_GAAG**[10]**[]**AAGA[14]**GA[2]AAGA**[4]**GAAA[1]**GA[4]**GAAA[1]GA[3]AAATGAAAGAAAAAGAA[1]**GGAA[6]GAAA[12]TAAA[1]**	12+4+3+3+15 10+14+1+4+2+6+12+1+4+3	71.2	71.2	This study
DYS1001_2	468.2	478	AGAA[1]AAGA**[4]**AGGA**[3]**AAGA[1]GAAA**[3]GGAA[1]GAAA[1]GGAA[10]** CE25_GAAG**[12]**[]**AAGA[15]GA[2]AAGA[4]**GAAA[1]**GA[4]**GAAA[1]GA[3]AAATGAAAGAAAAAGAA[1]GGAA**[6]GAAA[13]TAAA[1]**	+3+1+1+10 12+15+1+4+2+6+13+1+4+3	72	72	This study
DYS1001_3	487.9	498	AGAA[1]AAGA**[4]**AGGA**[3]**AAGA[1]GAAA**[3]GGAA[1]GAAA[1]GGAA[11]**	+3+1+1+11	77	77	This study
DYF403S1a_1	318.1	324	CE15_TTCT[2]TTCC[1]**TTCT[11]**ATC[1]**TTCT[15]**	11+15	26	11	Ballantyne et al. 2014
DYF403S1a_2	326.0	332	CE17_TTCT[2]TTCC[1]**TTCT[13]**ATC[1]**TTCT[15]**	13+15	28	13	Ballantyne et al. 2014
DYF403S1a_3	347.1	353	CE28_TTTC[2]GTTC[1]TCTT[2]TTTC[1]TTTT[1]**TCTT[17]TTTC[4]**	17+4	21	18.1	Ballantyne et al. 2014
DYS442	387.8	388	CE17_**GATA[12]GACA[3]**GATA[2]	15	15	15	This study
DYF403S1b	456.6	465	CE57_GAAA**[4]**AAGA**[15]**AGGA**[12]**AAGA**[9]**GAAA**[12]**AAGA[2]GAAC[1]GAAA[2]	4+15+12+9+12	52	51	Ballantyne et al. 2014

[#]Sequence blocks displayed in bold are used for the inferred repeat number

Supplementary Table S4: Assigning the repeat numbers to NA24385

Marker name	CE Length	MPS Length	STRnaming nomenclature[#]	Rationale	Inferred repeat number	Final repeat number	Source final repeat numbers
DYF393S1	182.4	190	CE26_GAA[4]AAG[24]	4+24	28	28	This study
DYS627	269.3	277	CE19_GA[3]CACA[1]GGGA[1]GAAA[3]GGAAAA[1]AAGA[3]GA[4]GAAA[19]AAGA[1]GAAA[1]GA[3]GGGA[2	3+3+19	25	21	YF Plus
DYS570	334.0	336	CE18_**TTTC**[18]	18	18	18	PPY23+YF Plus
DYS713	425.1	426	CE22_**TTTC**[19]TCTT[2]TCTG[1]TCTT[1]T[1]TTTC[2]**TCTT**[13]**TTTC**[5][]TTAT[1]**TATT**[4]	19+13+5+4	41	44	Zhang et al. 2012
DYS526b	483.2	485	CE38_CTTC[1]CCTC[3]TTTC[1]TTCC[1]TTTC[2]**TTTC**[12]CTTC[9]TTTC[1]CTTC[1]TTTC[1]CTTC[1]ATTT[1]CTTC[1]TTTC[2][]**TCCT**[12]	3+2+12+9+12	38	35	Ballantyne et al. 2014
DYF1000_1	248.3	249	CE60_AAG[9]CAG[1]AAG[23]AGG[1]AAG[1]AGG[1]AAG[24]	9+23+1+1+1+24	59	59	This study
DYF1000_2	251.3	252	CE61_AAG[9]CAG[1]AAG[20]AGG[1]AAG[1]AGG[1]AAG[1]AGG[1]AAG[26]	9+20+1+1+1+1+26	60	60	This study
DYF1000_3	280.0	281	CE32_AAG[5]AAGGAG[6]AAG[3]GAG[3]GA[1]GAA[6]GGA[2]AGA[5]AGG[1]AGA[1]GAA[2]G[1]GAA[4]GAT[1]GAA[2]AGA[24]	5+12+6+5+4+24	56	69.2	This study
DYF1000_4	289.0	290	CE34_AAG[5]AAGGAG[6]AAG[3]GAG[3]GA[1]GAA[6]GGA[2]AGA[5]AGG[1]AGA[1]GAA[2]G[1]GAA[4]GAT[1]GAA[2]AGA[27]	5+12+6+5+4+27	59	72.2	This study
DYS518	364.1	372	CE25_GAAA[2]GGAA[1]GAAA[16]GGGA[1]GAAA[4]GGAA[1]GA[1]GAAA[14]GGAA[1]GA[1]GGGA[2]AGAA[1]GAAA[1]AGAA[2]GGAA[3] CE36_AAAG[1]AAGA[2]AGGA[1]AAGA[3]AAGG[1]AACG[1]AAGG[11]AAGA[3]GAAA[3]AAAG[1]AGGG[3]AAGG[1]AGGA[3][]GA[4]	16+4+14+3	37	39	YF Plus
DYS1003	432.5	442	ACGG[1]AAGA[17]AAGG[1]TAGA[1]AAGA[9]GAAA[5]AAGA[3] CE32_AAGA[13]AAGG[1]AAAG[1]AA[1]AAGA[2]AAAG[3]AA[1]AAGA[5]AAAG[3]AAGG[1]AAAG[1]AAGG[1]AAAG[1]AAGG[1]AAAG[1]	3+11+3+3+3+3+17+9+5+3	60	60	This study
DYS1012	265.5	275	AAGG[1]AAAG[1]AAGG[2]AAAG[15]	13+3+5+3+15	39	39	This study
DYS1005	362.2	365	CE11.2_TTCCT[13]TCCC[7]TTCT[2]**CTTT**[5]**CCTT**[13]**CTTT**[10]CTT[1]CCTT[4]CTTT[2]	16.1+7+5+13+10+4	55.1	55.1	This study
DYS1010	490.8	499	CE20_AAGA[17]GA[15]AAGA[13]AAA[1]AAGA[1]GAGG[2]GA[3]	17+7.2+13	37.2	37.2	This study
DYS1007	268.8	273	CE35_AAGA[13]GAAA[3]GAGA[1]AAGA[4]GAAA[14][]GAAA[1]GA[5]CAGC[1]AAGA[1]GAAA[1]GA[5]	13+3+4+14	34	34	This study
DYR88_1	349.9	352	CE15_AAAG[15]	15	15	15	This study
DYR88_2	361.5	364	CE18_AAAG[18]	18	18	18	This study
DYF404S1_1	437.1	437	CE16_**TTTC**[16]TCTC[1]TTTC[1]	16	16	16	Ballantyne et al. 2014
DYF404S1_2	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
DYF387S1_1	508.7	508	CE38_A[7]GAA[4]AGAG[1]A[7]GAAA[2]GGTA[1]GGAA[4]GAAA[2]GGAA[1]GAAA[2]GGAA[9]GAAA[17]AT[1]A[8]	3+4+9+17	33	38	YF Plus
DYF387S1_2	516.7	516	CE40_A[7]GAA[4]AGAG[1]A[7]GAAA[2]GGTA[1]GGAA[4]GAAA[2]GGAA[1]GAAA[2]GGAA[10]GAAA[18]AT[1]A[8]	3+4+10+18	35	40	YF Plus
DYS1013	580.0	579	CE15_AGAT[12]TGAT[1]AGAT[1][]AGAT[2]AGAA[1]AGAT[6][]GATA[10][]AGAT[12]AGGC[3]	12+6+10+12+3	43	43	This study
DYS712	268.6	272	CE18_AGAT[12]AGAC[6]	12+6	18	18	Zhang et al. 2012
DYS711	362.9	364	CE31_TTC[4]T[1]**TTC**[29]TCC[3]TTC[1][]CTT[24]CTC[2]CTT[3]	4+29+24	57	61	Zhang et al. 2012
DYS626	478.8	488	CE18_AAAG[1]A[9]GCAG[1]AAAG[1]GAAG[1]A[5]GAAA[18][]GAAG[3]AAAG[3]AGACAG[1]AAAG[4]A[6]GAAA[2]GAAG[1]GAAA[3]	18+3+3+4+3	31	31	Ballantyne et al. 2014
DYF399S1_1	184.6	191	CE16_AAAG[3]AAAAAGA[1]AAAG[16]	3+16	19	20	Ballantyne et al. 2014
DYF399S1_2	188.5	195	CE17_AAAG[3]AAAAAGA[1]AAAG[16]_+2.1>AAAA_+19A>G	3+16+1	20	21	Ballantyne et al. 2014
DYF399S1_3	205.2	212	CE21.1_AAAG[3]AAAAAGA[1]AAAGG[1]AAAG[19]AAAA[1]	3+1.1+19+1	24.1	25.1	Ballantyne et al. 2014
DYS449	294.9	296	CE22_**TCTT**[11]TC[4]TCC[2]TCTT[2]TCCT[1]**TCTT**[3]TTCC[1]TCTT[1]TCCT[1]**TCTT**[16]	11+3+16	30	28	YF Plus
DYS724_1	368.5	375	CE16_CTTT[2]**TTTC**[4]TCTC[1]TTTC[2]TC[5]**CCTT**[13]T[1]CCTT[1]**CTTT**[19]	4+13+19	36	32	FTDNA
DYS724_2	388.0	395	CE21_CTTT[2]**TTTC**[4]TCTC[1]TTTC[2]TC[5]**CCTT**[16]T[1]CCTT[1]**CTTT**[21] CE21_**CTTC**[3]TTTC[1]**CTTC**[11]CTTT[1]**CTTC**[5]TTTC[1]CTCC[2][]CTCC[1]CTTC[1]CTCC[1]TTCC[1]CTCC[2]TTTC[1]T[5]**CTTT**[9]**CTAT**[1]	4+16+21	41	37	FTDNA
DYS547	479.7	481	**CTTT**[5]CCTTA[1]TTTC[1]TTCC[4]TTTC[1]**TCTT**[9]TTCT[2]CTCT[1]TTTC[3]T[4]CTTT[2]T[6]	3+11+5+9+1+5+4+9+3	50	46	Ballantyne et al. 2014
DYS576	169.9	178	CE18_AAAG[18]	18	18	18	PPY23+YF Plus
DYS612	271.1	274	CE37_CCT[5]CTT[1]**TCT**[4]CCT[1]**TCT**[26] CE38_GA[7]GAAA[4]TAGA[1]AAGA[14]GAAA[4]GGAA[1]AGAA[11]A[1]CAGAGA[1]AAGA[4]GAAA[13]AAGA[1]A[1]AAGA[2]GGAA[1]	5+4+26 3.2+4+14+4+11+0.1+	35	31	Forenseq
DYF1002_1	350.3	358	GAAA[4]AAGA[1]GAAA[2]GGAG[3]GGAA[2]AGAA[1]GGAA[1]GGAG[1] CE39_GA[7]GAAA[4]TAGA[1]AAGA[16]GAAA[4]GGAA[1]AGAA[12]A[1]CAGAGA[1]AAGA[4]GAAA[11]AAGA[1]A[1]AAGA[2]GGAA[1]	4+13+4+3 3.2+4+16+4+12+0.1+	60.3	60.3	This study
DYF1002_2	354.2	362	GAAA[4]AAGA[1]GAAA[2]GGAG[3]GGAA[2]AGAA[1]GGAA[1]GGAG[1 CE23_GAAG[12][]AAGA[16]GA[2]AAGA[4]GAAA[1]GA[2]GAAA[1]GA[3]AAATGAAAGAAAAAGAA[1]GGAA[6]GAAA[12]TAAA[1]AGAA[1]	4+11+4+3 12+16+1+4+1+6+12+1+4+3+	61.3	61.3	This study
DYS1001_1	479.8	490	AAGA[4]AGGA[3]AAGA[1]GAAA[3]GGAA[1]GAAA[1]GGAA[10] CE23_GAAG[10][]AAGA[15]GA[2]AAGA[4]GAAA[1]GA[4]GAAA[1]GA[3]AAATGAAAGAAAAAGAA[1]GGAA[6]GAAA[13]TAAA[1]AGAA[1]	3+1+1+10 10+15+1+4+2+6+13+1+4+3+	75	75	This study
DYS1001_2	479.8	490	AAGA[4]AGGA[3]AAGA[1]GAAA[3]GGAA[1]GAAA[1]GGAA[11] CE27.2_GAAG[15][]AAGA[14]GAAA[1]GA[3]AAGA[4]GAAA[1]GA[4]GAAA[1]GA[3]AAATGAAAGAAAAAGAA[1]GGAA[4]GAAA[11]AGAA[1]	3+1+1+11 15+14+1+1.2+4+2+4+11+4+	75	75	This study
DYS1001_3	498.1	508	AAGA[4]AGGA[3]AAGA[1]GAAA[3]GGAA[1]GAAA[1]GGAA[15]	3+3+1+1+15	79.2	79.2	This study
DYF403S1a_1	326.1	332	CE17_TTCT[2]TTCC[1]**TTCT**[14]ATC[1]**TTCT**[14]	14+14	28	13	Ballantyne et al. 2014
DYF403S1a_2	329.9	336	CE18_TTCT[2]TTCC[1]**TTCT**[14]ATC[1]**TTCT**[15]	14+15	29	14	Ballantyne et al. 2014
DYF403S1a_3	335.5	341	CE25_TTTC[2]GTTC[1]TCTT[2]TTTC[1]TTTT[1]**TCTT**[14]**TTTC**[4]	14+4	18	15.1	Ballantyne et al. 2014
DYS442	387.8	388	CE17_GATA[12]GACA[3]GATA[2]	12+3	15	15	This study
DYF403S1b	452.7	461	CE56_GAAA[4]AAGA[12]AGGA[11]AAGA[12]GAAA[12]AAGA[2]GAAC[1]GAAA[2]	4+12+11+12+12	51	50	Ballantyne et al. 2014

[#] Sequence blocks displayed in bold are used for the inferred repeat number

Supplementary Table S5: Assigning the repeat numbers to NA24631

Marker name	CE Length	MPS Length	STRnaming nomenclature[#]	Rationale	Inferred repeat number	Final repeat number	Source final repeat numbers
DYF399S1	182.5	190	CE26_GAA**[4]**AAG**[22]**CAG**[1]**AAG**[1]**	4+22+1+1	28	28	This study
DYS627	281	289	CE22_GA[3]CACA[1]GGGA[1]GAAA**[3]**GGAAAA[1]AAGA**[3]**GA[4]GAAA**[22]**AAGA[1]GAAA[1]GA[3]GGGA[2]	3+3+22	28	24	YF Plus
DYS570	346.1	348	CE21_**TTTC[21]**	21	21	21	PPY23+YF Plus
DYS713	417.1	418	CE20_**TTTC[16]**TCTT[2]TCTG[1]TCTT[1]T[1]TTTC[2]**TCTT[14]TTTC[5]**[]TTAT[1]**TATT[4]**	16+14+5+4	39	42	Zhang et al. 2012
DYS526b	483.1	485	CE38_CTTC[1]**CCTC[3]**TTTC[1]TTCC[1]TTTC[2]**TTCC[1]TTTC[11]CTTC[9]**TTTC[1]CTTC[1]TTTC[1]CTTC[1]ATTT[1]CTTC[1]TTTC[2][]**TCCT[14]**	3+1+11+9+14	38	35	Ballantyne et al. 2014
DYF1000_1	245.3	246	CE59_AAG**[9]**CAG[1]AAG**[49]**	9+49	58	58	This study
DYF1000_2	245.3	246	CE59_AAG**[9]**CAG[1]AAG**[27]**AGG**[1]**AAG**[21]**	9+27+1+21	58	58	This study
DYF1000_3	285.9	287	CE34_AAG**[5]**AAGGAG**[7]**AAG[3]GAG[3]GA[1]GAA**[7]**GGA[2]AGA**[5]**AGG[1]AGA[1]GAA[2]G[1]GAA**[4]**GAT[1]GAA[2]AGA**[23]**	5+7+14+5+4+23	58	71.2	This study
DYF1000_4	294.9	296	CE36_AAG**[5]**AAGGAG**[7]**AAG[3]GAG[3]GA[1]GAA**[6]**GGA[2]AGA**[5]**AGG[1]AGA[1]GAA[2]G[1]GAA**[4]**GAT[1]GAA[2]AGA**[27]**	5+7+14+5+4+27	61	74.2	This study
DYS518	348.5	356	CE21_GAAA[2]GGAA[1]GAAA**[15]**GGGA[1]GAAA**[4]**GGAA[1]GA[1]GAAA**[11]**GGAA[1]GA[1]GGGA[2]AGAA[1]GAAA[1]AGAA[2]GGAA**[3]**	15+4+11+3	33	35	YF Plus
			CE43_AAAG[1]AAGA[2]AGGA[1]AAGA**[3]**AAGG[1]AACG[1]AAGG**[10]**AAGA**[3]**GAAA**[3]**AAAG[1]AGGG**[3]**AAGG[1]AGGA**[3]**[]GA[4]	3+10+3+3+3+3+			
DYS1003	459.9	470	ACGG[1]AAGA**[19]**AAGG[1]TAGA[1]AAGA**[15]**GAAA**[5]**AAGA**[3]**	19+15+5+3	67	67	This study
			CE36_AAGA**[14]**AAGG[1]AAAG[1]AA[1]AAGA[2]AAAG**[3]**AA[1]AAGA**[5]**AAAG**[3]**AAGG[1]AAAG[1]AAGG[1]AAAG[1]AAGG[1]AAAG[1]				
DYS1012	281.2	291	AAGG[1]AAAG[1]AAGG[2]AAAG**[18]**	14+3+5+3+18	43	43	This study
DYS1005	362.4	365	CE11.2_**TTCCT[13]**TCCC**[6]**TTCT[2]**CTTT[5]CCTT[13]CTTT[11]**CTT[1]**CCTT[4]**CTTT[2]	16.1+6+5+13+11+4	55.1	55.1	This study
DYS1010	488.7	497	CE19.2_AAGA**[17]**GA**[12]**AAGA**[14]**AAA[1]AAGA[1]GAGG[2]GA[3]	17+6+14	37	37	This study
DYS1007	296	301	CE42_AAGA**[17]**GAAA**[5]**GAGA[1]AAGA**[5]**GAAA**[14]**[]GAAA[1]GA[5]CAGC[1]AAGA[1]GAAA[1]GA[5]	17+5+5+14	41	41	This study
DYR88_1	357.7	360	CE17_AAAG**[17]**	17	17	17	This study
DYR88_2	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
DYF404S1_1	429.1	429	CE14_**TTTC[14]**TCTC[1]TTTC[1]	14	14	14	Ballantyne et al. 2014
DYF404S1_2	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
DYF387S1_1	508.7	508	CE38_A[7]GAA**[4]**AGAG[1]A[7]GAAA[2]GGTA[1]GGAA**[4]**GAAA[2]GGAA[1]GAAA[2]GGAA**[10]**GAAA**[16]**AT[1]A[8]	3+4+10+16	33	38	YF Plus
DYF387S1_2	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
DYS1013	576	575	CE14_**AGAT[11]**TGAT[1]AGAT[1][]AGAT[2]AGAA[1]**AGAT[6]**[]**GATA[11]**[]**AGAT[11]AGGC[3]**	11+6+11+11+3	42	42	This study
DYS712	296.6	300	CE25_**AGAT[18]AGAC[7]**	18+7	25	25	Zhang et al. 2012
DYS711	368.8	370	CE33_**TTC[4]**T[1]**TTC[29]**TCC[3]TTC[1][]**CTT[26]**CTCCTC[1]CTT[3]	4+29+26	59	63	Zhang et al. 2012
DYS626	482.7	492	CE19_AAAG[1]A[9]GCAG[1]AAAG[1]GAAG[1]A[5]GAAA**[19]**[]GAAG**[3]**AAAG**[3]**AGACAG[1]AAAG**[4]**A[6]GAAA[2]GAAG[1]GAAA**[3]**	19+3+3+4+3	32	32	Ballantyne et al. 2014
DYF399S1_1	193.6	200	CE18.1_AAAG**[3]**AAAAAGA[1]AAAGG**[1]**AAAG**[17]**	3+1.1+17	22.1	22.1	Ballantyne et al. 2014
DYF399S1_2	197.7	204	CE19.1_AAAG**[3]**AAAAAGA[1]AAAGG**[1]**AAAG**[18]**	3+1.1+18	23.1	23.1	Ballantyne et al. 2014
DYF399S1_3	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
DYS449	310.8	312	CE26_**TCTT[12]**TC[4]TCC[2]TCTT[2]TCCT[1]**TCTT[3]**TTCC[1]TCTT[1]TCCT[1]**TCTT[19]**	12+3+19	34	32	YF Plus
DYS724_1	392.1	399	CE22_CTTT[2]**TTTC[4]**TCTC[1]TTTC[2]TC[5]**CCTT[19]**T[1]CCTT[1]**CTTT[19]**	4+19+19	42	38	FTDNA
DYS724_2	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
			CE24_**CTTC[3]TCCT[1]**TTTC[1]**CTTC[11]**CTTT[1]**CTTC[5]**TTTC[1]CTCC[2][]CTCC[1]CTTC[1]CTCC[1]TTCC[1]CTCC[2]TTTC[1]T[5]				
DYS547	491.9	493	**CTTT[13]**CCTTA[1]TTTC[1]**TTCC[4]**TTTC[1]**TCTT[13]**TTCT[2]CTCT[1]**TTTC[3]**T[4]CTTT[2]T[6]	3+1+11+5+13+4+13+3	53	49	Ballantyne et al. 2014
DYS576	166.1	174	CE17_AAAG**[17]**	17	17	17	PPY23+YF Plus
DYS612	274.1	277	CE38_**CCT[5]**CTT[1]**TCT[4]**CCT[1]**TCT[27]**	5+4+27	36	32	Forenseq
			CE33_GA**[7]**GAAA**[4]**TAGA[1]AAGA**[11]**GAAA**[4]**GGAA[1]AGAA**[11]A[1]**CAGAGA[1]AAGA**[4]**GAAA**[11]**AAGA[1]A[1]AAGA[2]GGAA[1]	3.2+4+11+4+11+			
DYF1002_1	330.7	338	GAAA**[4]**AAGA[1]GAAA[2]GGAG**[3]**GGAA[2]AGAA[1]GGAA[1]GGAG[1]	0.1+4+11+4+3	55.3	55.3	This study
			CE40_GA**[7]**GAAA**[4]**TAGA[1]AAGA**[17]**GAAA**[5]**GGAA[1]AGAA**[10]A[1]**CAGAGA[1]AAGA**[4]**GAAA**[12]**AAGA[1]A[1]AAGA[2]GGAA[1]	3.2+4+17+5+10+			
DYF1002_2	358.1	366	GAAA**[4]**AAGA[1]GAAA[2]GGAG**[3]**GGAA[2]AGAA[1]GGAA[1]GGAG[1]	0.1+4+12+4+3	62.3	62.3	This study
			CE23_GAAG**[12]**[]**AAGA[15]GA[2]AAGA[4]**GAAA[1]**GA[4]**GAAA[1]GA[3]AAATGAAAGAAAAAGAA[1]GGAA**[6]GAAA[12]TAAA[1]**	12+15+1+4+2+6+12+			
DYS1001_1	480	490	AGAA[1]**AAGA[4]AGGA[3]**AAGA[1]GAAA**[3]**GGAA[1]GAAA**[1]**GGAA**[10]**	1+4+3+3+1+1+10	75	75	This study
			CE24.2_GAAG**[12]**[]**AAGA[10]GAAA[1]GA[3]AAGA[4]**GAAA[1]**GA[4]**GAAA[1]GA[3]AAATGAAAGAAAAAGAA[1]GGAA**[4]GAAA[17]**	12+10+1+1.2+4+2+4+1			
DYS1001_2	486	496	AGAA[1]**AAGA[4]AGGA[3]**AAGA[1]GAAA**[3]GGAA[1]GAAA[1]GGAA[13]**	7+4+3+3+1+1+13	76.2	76.2	This study
DYS1001_3	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
DYF403S1a_1	322	328	CE16_TTCT[2]TTCC[1]**TTCT[14]**ATC[1]**TTCT[13]**	14+13	27	12	Ballantyne et al. 2014
DYF403S1a_2	333.9	340	CE19_TTCT[2]TTCC[1]**TTCT[15]**ATC[1]**TTCT[15]**	15+15	30	15	Ballantyne et al. 2014
DYF403S1a_3	339.4	345	CE26_TTTC[2]ATTC[1]TCTT[2]TTTC[1]TTTT[1]**TCTT[15]TTTC[4]**	15+4	19	16.1	Ballantyne et al. 2014
DYS442	387.8	429	CE17_**GATA[12]GACA[3]**GATA[2]	12+3	15	15	This study
DYF403S1b	421.6	388	CE48_GAAA**[4]**AAGA**[11]**AGGA**[9]**AAGA**[9]**GAAA**[10]**AAGA[2]GAAC[1]GAAA[2]	4+11+9+9+10	43	42	Ballantyne et al. 2014

[#] Sequence blocks displayed in bold are used for the inferred repeat number

Chapter 4

Improving the differentiation of closely related males by RMplex analysis of 30 Y-STRs with high mutation rates

Franz Neuhuber[1], Bettina Dunkelmann[1], Ines Grießner[1], Katharina Helm[1],

Manfred Kayser[2], Arwin Ralf[2]

[1] Department of Forensic Medicine and Forensic Neuropsychiatry, University of Salzburg, Salzburg, Austria
[2] Department of Genetic Identification, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands

Published in *Forensic Science International: Genetics*, May 2022
Volume 58, 102682
doi: https://doi.org/10.1016/j.fsigen.2022.102682

Abstract

The discovery of rapidly mutating (RM) Y-STRs started to move the field of forensic Y-STR analysis from male lineage identification towards male individual identification. Previously, the forensic value of RM Y-STRs for differentiating male relatives was limited due to the modest number of 13 identified RM Y-STRs. Recently, new RM Y-STRs were discovered, with strong expectations for significantly improving male relative differentiation; however, empirical evidence is missing yet. More recently, the genotyping method RMplex for efficiently analyzing 30 Y-STRs with high mutation rates, including all 26 currently known RM Y-STRs, was introduced. Here, we applied RMplex as well as the current state-of-the-art commercial Y-STR kit: Yfiler™ Plus PCR Amplification kit, to several hundreds of DNA-confirmed father-son pairs. Newly established estimates confirmed the high mutation rates of novel and previous RM Y-STRs. By combining current with previous data, we provide updated consensus estimates of mutation rates for all 49 Y-STRs targeted with both methods. Based on RMplex, 42% of 499 father-son pairs were differentiated, while 14% of 530 pairs based on Yfiler™ Plus, and 48% of 499 pairs based on both methods combined. Regarding brothers, RMplex also clearly outperformed Yfiler Plus, with differentiation rates of 62% and 33%, respectively. By combining both methods 72.9% of the brothers showed at least one mutation. For unrelated males, both methods achieved a discrimination capacity of 99.8% and a haplotype diversity of 0.999991, since all males had different haplotypes, except for two, perhaps indicating a hidden paternal relationship. Overall, this study underlines the value of RM Y-STRs in general and RMplex in particular for differentiating male relatives highly relevant in forensic genetics. It provides the first empirical evidence on the high value of RMplex for differentiating close male relatives, which for father-son pairs was almost 60% higher than with the initial set of 13 RM Y-STRs and three times higher than with Yfiler™ Plus. Based on our results from closely related males, we expect RMplex to also improve the differentiation of more distantly related males significantly, which needs empirical demonstration in future studies. We encourage the forensic community to apply RMplex in all forensic cases where a match with a commercial Y-STR kit was obtained between the male suspect and the evidence material, or to solely use RMplex in such cases, aiming to find out if the male suspect or any of his male paternal relatives left the evidence material at the crime scene.

Introduction

Male-specific Y-chromosomal short tandem repeats (Y-STRs) have several applications in forensic genetics: they are used to predict the paternal ancestry of an unknown donor of a crime scene trace, to infer surnames, or for familial searching to trace unknown perpetrators. Until now, however, the most valuable contribution of Y-STRs to the forensic toolbox is their application in unbalanced male-female mixtures, as are often observed in sexual assault cases, where typically standard autosomal STR-profiling is not informative for identifying the male perpetrator. In such cases, Y-STRs are applied to characterize the paternal lineage of the male contributor aiming to identify the male perpetrator (1). The largest limitation of this application is that typically Y-STR haplotypes are shared between many paternally related (2, 3) and sometimes even unrelated males (4-6). This poses a problem in how to evaluate in court the evidentiary value of a match between the suspect's Y-STR profile (haplotype) and that from the crime scene material, as the sample donor could be the suspect or, with the very same probability, one of his male paternal relatives (7-11). The number of haplotype matches of unrelated individuals can be significantly reduced by increasing the number of Y-STRs analyzed (12, 13). Nevertheless, the problem of haplotype sharing between many paternally related males remains.

Rapidly mutating Y-STRs (RM Y-STRs) have been proposed as a solution to overcome this problem (14). Their increased mutation rates (i.e., one or a few mutations per locus per 100 generations) allow to differentiate more paternally related males than possible with Y-STRs characterized by moderate mutation rates (i.e., one or a few mutations per locus per 1000 generations). As haplotypes derived from RM Y-STRs are shared by less relatives compared to haplotypes based on Y-STRs with moderate mutation rates, the evidentiary value of a haplotype match based on RM Y-STRs is higher. However, until recently, the number of known RM Y-STRs was limited to the 13 initially discovered markers (15) and thus was the male relative differentiations rates they can achieve, i.e., 27% for father-sons (one meiosis), 46% for brothers and grandfather-grandson pairs (two meioses), and 62% for cousins (four meioses) (2).

To help facilitating the more efficient differentiation of unrelated and related males available with RM Y-STRs (4), the current generations of widely used commercial Y-STR kits, such as Yfiler™ Plus PCR Amplification kit (Thermo Fisher Scientific) (in the following referred to as Yfiler Plus) or PowerPlex® Y23 System (Promega) include some of the 13 initially identified RM Y-STRs (six and two, respectively). The actual male relative differentiation rates obtainable with such latest generation commercial Y-STR kits has yet to be empirically determined. They are expected to be lower than that obtained for the

full set of 13 RM Y-STRs, because most of the 13 markers, including those with the highest mutation rates, were not included in these state-of-the-art commercial Y-STR kits. Nevertheless, the relative differentiation rates obtained with these commercial kits will be higher than those of their predecessors: AmpFLSTR™ Yfiler™ PCR Amplification Kit (in the following referred to as Yfiler), previously established at 5% for father-sons (2), and PowerPlex® Y System, because both previous kits include fewer Y-STR loci and no RM Y-STRs.

Recently, 12 novel RM Y-STRs were identified (16), almost doubling the number of available RM Y-STRs. Empirical male relative differentiation rates obtainable with the full set of currently known 26 RM Y-STRs are not available as of yet. Theoretically expected male differentiation rates previously calculated based on the locus-specific mutation rates inform that this increased RM Y-STR set should allow differentiating father and sons with a rate of approximately 44%, brothers with 69% and cousins (separated by four meioses) with 90% (16). More recently, a novel genotyping method termed RMplex (17) was introduced for the effective analysis of 30 Y-STRs with high mutation rates, including all currently known 26 RM Y-STRs together with four additional Y-STRs with elevated mutation rates.

In the present study, we apply RMplex for analyzing 30 Y-STRs with high mutations rates to the closest possible, and therefore most challenging to differentiate, paternal male relatives by analyzing 530 DNA-confirmed father-son pairs to empirically quantify mutation rates as well as differentiation rates of father-sons, brothers, and unrelated males. Moreover, we compare these findings to those we obtained with the current state-of-the-art commercial Y-STR kit Yfiler Plus that targets 25 Y-STRs including six RM Y-STRs that overlap with RMplex. Our study is the first to deliver i) independent confirmation of mutation rates for the recently discovered novel RM Y-STRs, ii) empirical father-son and brother differentiation rates for RMplex and Yfiler Plus, and iii) empirical differentiation of unrelated males with RMplex and Yfiler Plus. Moreover, we provide updated consensus mutation rate estimates for 49 Y-STR included in both methods by combining the data produced in the present study with those from the literature.

Material and methods

DNA samples

A total of 1016 male individuals from Austria with first degree relationships to at least one other male individual in the study were analyzed including 482 fathers with a total of 534

sons. In the cases where a father had more than one son included, each of the sons formed a pair to be analyzed with the same father. Considering father-son pairs with the same father (but different sons) may cause some dependency as the alleles of the father determine the mutations in the son. However, the number of dependent pairs with the same father was with 73 relatively small compared to the number of independent father-son pairs (457); hence no large impact on the results is expected. Amongst the sons were 4 pairs of monozygotic twins of which only one was considered for the consequent analyses and the other was excluded. Moreover, the sample set contained two dizygotic twin pairs, which were considered as brothers. In total, 530 father-son pairs and 92 brother pairs could be defined in the sample set.

All DNA samples come from previous paternity testing using the full mother-father-son/sons settings, where achieved paternity probabilities based on autosomal STRs were >99.9999%. This allows concluding true biological father-son relationships for all of the pairs included in this study, which serves as key prerequisite for mutation rate and father-son differentiation rate analysis. The samples were fully anonymized and only the father-son relationships (and the brother relationships by extension) were conserved. The ethics committee of the University of Salzburg approved this study (EK-GZ: 07/2021).

Genotyping

RMplex is a genotyping assay that consists of two non-overlapping sets of Y-STRs of which the results are analyzed combined. The total of 30 Y-STRs contain various multi-copy markers leading to a total of 44 loci being amplified. Genomic DNA was extracted using the chelex 100 extraction method. The exact concentrations were unknown; however, these samples were previously used for paternity testing, here the same amounts of DNA were used. PCR amplifications were performed under the same conditions as described previously (17). For DYS570 the alternative forward primer that was suggested in the original method publication (17) was used. The only deviation from the original protocol was that here the amplification was performed in a reduced reaction volume of 10 µL. The DNA was amplified in PE 9700 Thermal Cycler (Perkin Elmer). The resulting amplification products were separated using a 3500 Genetic Analyzer (Thermo Fisher Scientific); and the resulting electropherograms were analyzed using Genemapper ID-X, Version 1.4. For Yfiler Plus, the same instruments were used, the manufacturer's recommendations were followed, except for the PCR reaction volume which was reduced to 12.5 µL. For both methods, allele calling was done by two experienced analysts independently, and in case of conflicting outcomes a third expert was involved for clarification.

Data analysis

Mutation rates and male relative differentiation rates were calculated using the frequentist approach. On top of the data generated here, mutations rate data was also obtained from literature, for a fair comparison, the reference mutation rates were recalculated using the frequentist approach in case the original publication used a different (i.e., Bayesian) approach. The 95% confidence intervals of the mutation and male relative differentiation rates were calculated using the Clopper–Pearson interval (18). Different mutation rates estimates were compared to each other using Fisher's exact tests (19); Bonferroni correction (20) was applied to account for multiple testing. For the mutation rate estimates, all pairs were used for analysis, in case of partial profiles the pair was excluded for the analysis of only the specific Y-STRs that had missing data. For the male relative differentiation rate and the haplotype diversity estimations, pairs and individuals, respectively, with missing data for a single or more Y-STRs were excluded from the analysis.

Results and Discussion

Assay performances

All DNA samples were genotyped using three assays: the two multiplex assays of the non-commercial RMplex and the single multiplex assay of the commercial Yfiler Plus kit. While Yfiler Plus delivered complete Y-STR profiles for all 1016 individual DNA samples from all 530 father-son pairs, RMplex achieved complete Y-STR profiles for 983 (97%) individual samples from 499 (94%) of the 530 pairs. A total of 31 (3.1%) out of the 1016 samples showed at least one locus dropout with RMplex 1, while with RMplex 2 locus-dropouts were seen in six (0.6%) samples. Locus-dropouts with RMplex were mostly caused by three Y-STRs i.e., DYR88, DYS1013, and DYF404S1, which delivered no result in 19 (1.9%), 13 (1.3%) and 9 (0.9%) of the 1016 individual DNA samples, respectively. Notably, these same Y-STRs also showed a reduced performance in the initial RMplex validation study (17). Overall, the technical performance of Yfiler Plus was superior to that of RMplex this may in part be explained by RMplex being less sensitive than Yfiler Plus (17, 21). Another reason may be that RMplex contains Y-STRs with larger PCR fragment sizes than the longest ones of Yfiler Plus and therefore is more prone to reduced performance caused by low DNA quantity and/or quality. Lastly, Yfiler Plus may be more resilient against PCR inhibitors. In general, it is not unexpected to see that genotyping assays developed by

academia, such as RMplex, show reduced performance compared to those developed by industry, such as Yfiler Plus, as commercial companies typically spend more resources on the assay development than are available to academia. Of the six Y-STRs that overlap between Yfiler Plus and RMplex, all obtained results were fully concordant between these two genotyping methods.

Mutation analysis

In total, we identified 333 Y-STR mutations among the 530 father-son pairs analyzed, of which 289 were detected with RMplex and 76 with Yfiler Plus including 32 mutations at the six RM Y-STRs targeted by both methods. Among the 333 Y-STR mutations observed, 325 were single-step repeat mutations (97%) and 9 were multi-step repeat mutations (3%). These results, showing a vast excess of single-step over multi-step repeat mutations, are in line with those from previous studies of Y-STR mutations in father-son pairs (15, 16). Moreover, we observed slightly more repeat contractions: 175 (53%) than repeat expansions: 158 (44%); but overall, the number of contractions and expansions were very similar, which also agrees with previous findings (15, 16). All observed mutations with the genotype of both father and son are shown in supplementary Table S1.

On average, RMplex detected a mean number of 0.54 mutations per pair with a standard deviation of 0.73 and a range of 0 to 3 mutations per pair. In contrast, Yfiler Plus detected a mean number of 0.14 mutations per father-son pair with a standard deviation of 0.38 and a range from 0 to 3 mutations per pair. When combining both methods and using all 49 Y-STRs together, a mean number of 0.69 mutations per pair was found with a standard deviation of 0.89 and ranging from 0 to 4 mutations per pair.

We estimated locus-specific mutation rates for all 49 Y-STRs targeted with both methods (Table 1) and compared these newly estimated mutations rates with those previously reported in father-son based studies for the same Y-STRs (15, 16). Based on Fisher's exact test (see Table 1 for the p-values), three of the 49 Y-STRs showed p-values below the 0.05 nominal significance threshold, which were DYS547 (p-value 0.0093), DYS1012 (p-value 0.0276), and DYS518 (p-value 0.0180). For these three Y-STRs, the mutation rate estimates in the present study were lower than described in the reference literature. However, when correcting for multiple testing using Bonferroni correction with an adjusted significance threshold of 0.0010, none of the Y-STRs showed a significant mutation rate difference. We also found Y-STRs displaying higher mutation rate estimates than described in literature, e.g., DYS712, DYS1007, and DYS449, albeit none of those showed a statistically significant difference.

Table 1: Empirically established locus-specific mutation rates obtained for 49 Y-STRs based on RMplex and Yfiler Plus data from a total of 530 DNA-confirmed father-son pairs.

Marker	Assay	Total pairs	Mutations	Expansions	Contractions	Mutation rate ($x10^{-3}$)	95% confidence interval ($x10^{-3}$)	Reference mutation rate ($x10^{-3}$) #	p-value
DYF399S1	RMplex	530	41	14	27	77.4	56.1-103.5	77.5	1.0000
DYS724	RMplex	529	28	14	14	52.9	35.5-75.6	46.4	0.5583
DYS712	RMplex	530	23	8	15	43.4	27.7-64.4	27.2	0.0852
DYF1001	RMplex	528	19	13	6	36.0	21.8-55.6	52.0	0.1593
DYF1000	RMplex	530	19	15	4	35.8	21.7-55.4	35.9	1.0000
DYF403S1a	RMplex	530	15	8	7	28.3	15.9-46.3	30.6	0.8828
DYS711	RMplex	530	14	5	9	26.4	14.5-43.9	26.6	1.0000
DYS1007	RMplex	530	12	6	6	22.6	11.8-39.2	15.5	0.2553
DYS449	Yfiler Plus+RMplex	530	11	9	2	20.8	10.4-36.8	11.8	0.1363
DYR88	RMplex	511	9	3	6	17.6	8.1-33.2	29.1	0.2039
DYS713	RMplex	529	9	4	5	17.0	7.8-32.0	14.2	0.6797
DYS612	RMplex	530	8	3	5	15.1	6.5-29.5	14.1	0.8366
DYS1013	RMplex	517	7	5	2	13.5	5.5-27.7	9.9	0.4683
DYS1010	RMplex	527	7	2	5	13.3	5.4-27.2	14.2	1.0000
DYS526b	RMplex	528	7	3	4	13.3	5.3-27.1	12.1	0.8225
DYS458	Yfiler Plus	530	7	3	4	13.2	5.3-27.0	8.0	0.2975
DYF1002	RMplex	530	7	3	4	13.2	5.3-27.0	17.9	0.5614
DYF403S1b	RMplex	529	6	2	4	11.3	4.2-24.5	11.4	1.0000
DYS570	Yfiler Plus+RMplex	530	6	2	4	11.3	4.2-24.5	11.9	1.0000
DYS385	Yfiler Plus	530	6	1	5	11.3	4.2-24.5	5.1	0.1286
DYF387S1	Yfiler Plus+RMplex	530	6	3	3	11.3	4.2-24.5	15.5	0.6795
DYS1003	RMplex	530	6	2	4	11.3	4.2-24.5	13.0	1.0000
DYS1005	RMplex	530	6	3	3	11.3	4.2-24.5	9.3	0.6195
DYS576	Yfiler Plus+RMplex	530	5	3	2	9.4	3.1-21.9	13.9	0.5142
DYS533	Yfiler Plus	530	5	2	3	9.4	3.1-21.9	4.6	0.1988
DYS460	Yfiler Plus	530	4	2	2	7.5	2.1-19.2	5.8	0.7512
DYF404S1	RMplex	520	3	2	1	5.8	1.2-16.8	12.1	0.4815
DYS547	RMplex	529	3	2	1	5.7	1.2-16.5	23.2	**0.0093**
DYS456	Yfiler Plus	530	3	2	1	5.7	1.2-16.5	4.6	0.7242
DYS393	Yfiler Plus	530	3	2	1	5.7	1.2-16.5	1.7	0.1424
DYS439	Yfiler Plus	530	3	1	2	5.7	1.2-16.5	3.5	0.4446
DYS481	Yfiler Plus	530	3	2	1	5.7	1.2-16.5	4.6	0.7253
DYS1012	RMplex	530	3	2	1	5.7	1.2-16.5	19.2	**0.0276**
DYS626	RMplex	529	2	1	1	3.8	0.5-13.6	11.8	0.1317
DYS627	Yfiler Plus+RMplex	530	2	2	0	3.8	0.5-13.6	11.9	0.1342
DYS19	Yfiler Plus	530	2	0	2	3.8	0.5-13.6	4.0	1.0000
YGATAH4	Yfiler Plus	530	2	1	1	3.8	0.5-13.6	2.8	0.6665
DYS391	Yfiler Plus	530	2	1	1	3.8	0.5-13.6	2.8	0.6660
DYS518	Yfiler Plus+RMplex	530	2	1	1	3.8	0.5-13.6	18.0	**0.0180**
DYS437	Yfiler Plus	530	2	0	2	3.8	0.5-13.6	1.1	0.2307
DYF393S1	RMplex	530	2	0	2	3.8	0.5-13.6	8.2	0.3869
DYS442	RMplex	528	1	0	1	1.9	0-10.5	9.4	0.1356
DYS389I	Yfiler Plus	530	1	1	0	1.9	0-10.5	5.1	0.4695
DYS448	Yfiler Plus	530	1	0	1	1.9	0-10.5	0.0	0.2328
DYS635	Yfiler Plus	530	0	0	0	0.0	0-6.9	3.5	0.3463
DYS389II	Yfiler Plus	530	0	0	0	0.0	0-6.9	3.4	0.3464
DYS390	Yfiler Plus	530	0	0	0	0.0	0-6.9	1.1	1.0000
DYS438	Yfiler Plus	530	0	0	0	0.0	0-6.9	0.6	1.0000
DYS392	Yfiler Plus	530	0	0	0	0.0	0-6.9	0.6	1.0000

Reference mutation rates were those combined from Ballantyne et al. 2010 (15) and Ralf et al. 2020 (16).

These observed mutation rate differences, none being statistically significant after multiple testing correction, may be related to stochastic effects caused by small sample size, given that Y-STR mutations being relatively rare events even for those with increased mutation rates. Since approximately three fold more father-son pairs were included in the previous mutation rate studies (15, 16) compared to the present one, it may be expected that the previously obtained mutation rate estimates are closer to the ground truth than those reported here. Nevertheless, it should be noted that mutation rates estimates can show clear differences between studies; hence, the most conservative approach may be to combine data from multiple studies.

To this end, we carried out an extensive, yet not exhaustive, literature search for published Y-STR mutation data based on father-son pair analysis involving the 49 Y-STRs targeted here with both methods. We pooled the data of the present study with those obtained from 31 previous studies (2-4, 15, 16, 22-47) for the same loci (supplementary Table S2), covering a total ranging from 2,025 to 12,387 father-son pairs depending on the Y-STR marker. The newly established updated locus-specific consensus mutation rate estimates are presented in Table 2, and could serve as a new reference for future studies.

Moreover, using these updated consensus mutation rates, we revisited the four-category classification system that we previously proposed (16) to classify the 49 Y-STRs (Table 2). Twenty-four Y-STRs were classified as RM Y-STRs (mutation rates >1 x 10^{-2}), of which previously 23 were described as such (15, 16) and one (DYS1013) as fast mutating (FM) Y-STR (16). Nine Y-STRs were classified as FM Y-STRs (mutation rates 5 x 10^{-3} – 1 x 10^{-2}), of which previously five were previously descripted as such, three as RM Y-STRs (DYS403S1b, DYS626, and DYS570) (15), and one (DYS389II) as moderately mutating (MM) Y-STR. Thirteen Y-STRs were classified as MM Y-STRs (mutation rate 1 x 10^{-3} – 5 x 10^{-3}), of which previously 11 were described as such and two DYS460 and DYS389I) as FM- Y-STRs (15). The remaining three Y-STRs (DYS448, DYS392, and DYS438) were classified as slowly mutating (SM) Y-STRs (mutation rate <10^{-3}) and were previously described as such (15). According to this revisited classification, RMplex contains 24 RM Y-STRs (80%) and six FM Y-STRs (20%), whereas Yfiler Plus includes five RM Y-STRs (20%), four FM Y-STRs (16%), 13 MM Y-STRs (52%) and three SM Y-STRs (12%).

The differences between the current classification and that reported previously may reflect uncertainties in the mutation rate estimates and the difficulty of categorizing Y-STRs by using sharp mutation rate borders. For Y-STRs with mutation rates close to the defined borders, a slight increase of the updated mutation rate results in a classification upgrade (e.g. DYS1013), while a slight decrease in a downgrade (e.g. DYS1005). However, by increasing the sample size of the mutation rate underlying father-son pairs further and

further, the consensus estimates will become more and more robust, decreasing fluctuations in the next updated mutation rate estimates and thus classification changes. As previously emphasized (16), it is our opinion that the four-category classification system does provide a practically useful way to group (Y-)STRs based on their mutability.

In general, differences in mutation rates between studies done in different populations could also reflect biological differences of these populations, which e.g., can be linked to differences in haplogroup compositions of the populations (48). To determine if such effects could be seen for the 49 Y-STRs that were analyzed in the current study, we established separate datasets of father-son pair based mutation data for the two major populations for which most mutation data were available in the literature and including the present data, namely for Europeans and for Asians (Table 2). For the most recently discovered Y-STRs, mutation rate data from Asian populations are not yet available; hence, these markers had to be excluded from this analysis, leaving a total of 35 Y-STRs in this analysis. Comparing mutation rate estimates obtained from Europeans and Asians using Fisher's exact test revealed three of the 35 Y-STRs tested with p-values below the 5% nominal significance threshold: DYF399S1 (p-value 0.0011), DYS570 (p-value 0.0177), and DYS19 (p-value 0.0417). For all three markers, the mutation rate estimates based on the European males were higher than those from Asian males. However, after Bonferroni correction for multiple testing, none of the differences remained statistically significant. Despite not being statistically significant, we noted Y-STRs with over two-fold differences in mutation rate estimates between these two major populations, e.g., DYS390, DYS389I, DYS19, and DYS392 (Table 2), which are all Y-STRs with lower mutation rates. It is not surprising that larger differences are most common in Y-STRs with lower mutation rates, where due to low numbers of mutational events observed, stochastic effects have a large impact. Notably, the observed mutation rate differences could lead to different classifications of several Y-STRs, e.g., DYF387S1, DYS403S1b, DYS570 would be classified as RM Y-STRs based on European data, while based on the Asian data they would be FM Y-STRs. More data for different major populations are needed to get better insights into population effects on Y-STR mutation rates. If in the future, statistically significant differences in Y-STR mutation rates between major populations based on large-enough sample size would be established, it may be appropriate to use population-specific mutation rate estimates in future applications.

Table 2: Updated consensus estimates of locus-specific mutation rates for 49 Y-STRs by combining current data with literature data from father-son pair analyses.

	Overall consensus*					Europe			Asia			
Marker	Total pairs	Mutations	Mutation rate ($x10^{-3}$)	95% C.I. ($x10^{-3}$)	Classification	Total pairs	Mutations	Mutation rate ($x10^{-3}$)	Total pairs	Mutations	Mutation rate ($x10^{-3}$)	p-value
DYF399S1	7655	481	62.8	57.5 - 68.5	RM Y-STR	2324	180	77.5	4320	244	56.5	**0.0011**
DYF1001	2144	103	48.0	39.4 - 58.0	RM Y-STR	2144	103	48.0	n.a.	n.a.	n.a.	n.a.
DYS724	2145	103	48.0	39.4 - 57.9	RM Y-STR	2145	103	48.0	n.a.	n.a.	n.a.	n.a.
DYF1000	2146	77	35.9	28.4 - 44.6	RM Y-STR	2146	77	35.9	n.a.	n.a.	n.a.	n.a.
DYS712	2476	77	31.1	24.6 - 38.7	RM Y-STR	2146	67	31.2	330	10	30.3	1.0000
DYF403S1a	7265	198	27.3	23.6 - 31.3	RM Y-STR	2034	61	30.0	4320	117	27.1	0.5151
DYS711	2146	57	26.6	20.2 - 34.3	RM Y-STR	2146	57	26.6	n.a.	n.a.	n.a.	n.a.
DYR88	2127	56	26.3	19.9 - 34.1	RM Y-STR	2127	56	26.3	n.a.	n.a.	n.a.	n.a.
DYS1007	2146	37	17.2	12.2 - 23.7	RM Y-STR	2146	37	17.2	n.a.	n.a.	n.a.	n.a.
DYF1002	2146	36	16.8	11.8 - 23.1	RM Y-STR	2146	36	16.8	n.a.	n.a.	n.a.	n.a.
DYS612	8360	136	16.3	13.7 - 19.2	RM Y-STR	2668	39	14.6	4320	71	16.4	0.6211
DYS1012	2146	34	15.8	11.0 - 22.1	RM Y-STR	2146	34	15.8	n.a.	n.a.	n.a.	n.a.
DYS547	7900	116	14.7	12.1 - 17.6	RM Y-STR	2208	42	19.0	4320	57	13.2	0.0861

Table 2 (continued)

DYS627	11871	172	14.5	12.4 - 16.8	RM Y-STR	2667	35	13.1	7832	115	14.7	0.6367
DYS1010	2143	30	14.0	9.5 - 19.9	RM Y-STR	2143	30	14.0	n.a.	n.a.	n.a.	n.a.
DYS713	3752	52	13.9	10.4 - 18.1	RM Y-STR	2145	32	14.9	1607	20	12.4	0.5741
DYS518	11288	150	13.3	11.3 - 15.6	RM Y-STR	2086	30	14.4	7830	91	11.6	0.3126
DYS576	12387	157	12.7	10.8 - 14.8	RM Y-STR	2897	35	12.1	7762	101	13.0	0.7731
DYS1003	2146	27	12.6	8.3 - 18.3	RM Y-STR	2146	27	12.6	n.a.	n.a.	n.a.	n.a.
DYF404S1	8312	104	12.5	10.2 - 15.1	RM Y-STR	2259	24	10.6	5042	65	12.9	0.4889
DYS526b	7871	97	12.3	10.0 - 15.0	RM Y-STR	2179	27	12.4	4320	52	12.0	0.9049
DYS449	12303	138	11.2	9.4 - 13.2	RM Y-STR	2518	37	14.7	8413	85	10.1	0.0653
DYS1013	2133	23	10.8	6.8 - 16.1	RM Y-STR	2133	23	10.8	n.a.	n.a.	n.a.	n.a.
DYF387S1	11150	114	10.2	8.4 - 12.3	RM Y-STR	2334	34	14.6	7805	75	9.6	0.0510
DYS1005	2146	21	9.8	6.1 - 14.9	FM Y-STR	2146	21	9.8	n.a.	n.a.	n.a.	n.a.
DYF403S1b	7162	65	9.1	7.0 - 11.6	FM Y-STR	1931	22	11.4	4320	32	7.4	0.1381
DYS626	7910	68	8.6	6.7 - 10.9	FM Y-STR	2218	22	9.9	4320	35	8.1	0.4831
DYS458	11830	101	8.5	7.0 - 10.4	FM Y-STR	2555	22	8.6	7256	57	7.9	0.7005
DYS570	11717	97	8.3	6.7 - 10.1	FM Y-STR	2225	26	11.7	7764	50	6.4	**0.0177**
DYS385	11699	88	7.5	6.0 - 9.3	FM Y-STR	2561	15	5.9	7269	55	7.6	0.4153
DYS442	2025	15	7.4	4.2 - 12.2	FM Y-STR	2025	15	7.4	n.a.	n.a.	n.a.	n.a.

Table 2 (continued)

DYF393S1	2242	16	7.1	4.1 - 11.6	FM Y-STR	2242	16	7.1	n.a.	n.a.	n.a.	n.a.
DYS389II	11685	64	5.5	4.2 - 7.0	FM Y-STR	2542	8	3.1	7274	43	5.9	0.1091
DYS439	11687	56	4.8	3.6 - 6.2	MM Y-STR	2535	11	4.3	7282	35	4.8	0.8669
DYS481	8257	39	4.7	3.4 - 6.5	MM Y-STR	2543	12	4.7	5358	22	4.1	0.7145
DYS456	11488	50	4.4	3.2 - 5.7	MM Y-STR	2556	14	5.5	7274	27	3.7	0.2830
DYS460	8400	36	4.3	3.0 - 5.9	MM Y-STR	2887	15	5.2	5363	21	3.9	0.3879
DYS635	11472	44	3.8	2.8 - 5.1	MM Y-STR	2531	7	2.8	7283	32	4.4	0.3584
DYS533	8250	29	3.5	2.4 - 5.0	MM Y-STR	2529	14	5.5	5365	15	2.8	0.0723
DYS390	11070	30	2.7	1.8 - 3.9	MM Y-STR	2557	2	0.8	6644	19	2.9	0.0846
DYS391	11711	29	2.5	1.7 - 3.6	MM Y-STR	2558	9	3.5	7284	15	2.1	0.2416
DYS389I	11702	28	2.4	1.6 - 3.5	MM Y-STR	2550	10	3.9	7283	14	1.9	0.0999
DYS19	11707	23	2.0	1.2 - 2.9	MM Y-STR	2555	10	3.9	7283	11	1.5	**0.0417**
YGATAH4	11493	22	1.9	1.2 - 2.9	MM Y-STR	2554	8	3.1	7281	13	1.8	0.2154
DYS393	11702	20	1.7	1.0 - 2.6	MM Y-STR	2549	6	2.4	7284	10	1.4	0.3895
DYS437	11707	14	1.2	0.7 - 2	MM Y-STR	2559	4	1.6	7279	6	0.8	0.2972
DYS448	10735	9	0.8	0.4 - 1.6	SM Y-STR	2546	1	0.4	6531	4	0.6	1.0000
DYS392	11677	9	0.8	0.4 - 1.5	SM Y-STR	2527	1	0.4	7281	7	1.0	0.6888
DYS438	11702	3	0.3	0.1 - 0.7	SM Y-STR	2550	1	0.4	7283	2	0.3	1.0000

* Data from 32 previous studies [2-4, 15, 16, 19-43] were combined with current data, Table S2 shows the data per study.

Differentiating related males

Next, we used the data to derive empirical father-son differentiation rates, defined as percentage of father-son pairs out of all pairs analyzed that differ by at least one Y-STR mutation. Overall, RMplex yielded a markedly higher father-son differentiation rate than Yfiler Plus did and considering all Y-STRs from both methods combined led to a further, albeit more modest, increase compared to RMplex.

With Yfiler Plus, 71 of 530 father-son pairs were differentiated, resulting in a father-son differentiation rate of 13.4% (Figure 1). This differentiation rate is more than twice as high as previously obtained for the predecessor Yfiler which was 5.0% (2). The increased differentiation rate of Yfiler Plus compared to Yfiler was expected because 10 more loci, including six RM Y-STRs, were included in Yfiler Plus, hence increased chance for mutations to be observed. However, the differentiation rate of Yfiler Plus was still lower than obtained with the 13 initially identified RM Y-STRs such as previously reported as 26.5% (2). This is not unexpected because only six of the 13 RM Y-STRs are included in Yfiler Plus and because the effect of RM Y-STRs on father-son differentiation is stronger than that of Y-STRs with moderate mutation rates which Yfiler Plus consists of mostly.

Importantly, RMplex far outperformed Yfiler Plus with an estimated father-son differentiation rate of 41.9% (Figure 1), with 209 out of 499 pairs being differentiated. The father-son differentiation rate obtained with RMplex was three times higher than that obtained with Yfiler Plus and almost 60% higher than with the initial set of 13 RM Y-STRs. Combining the data from RMplex and Yfiler Plus and considering the total of 49 Y-STRs, a further increase of the differentiation rate to 47.5% (237 of 499 pairs) was noted (Figure 1). This increase of 5.6 percentage points compared to RMplex alone is rather small when considering that a relatively large number of 19 Y-STRs in Yfiler Plus (not overlapping with RMplex) was responsible for this. This reemphasizes once again that Y-STRs with moderate mutation rates can improve male relative differentiation, especially when being applied in larger numbers, but their effect is much smaller than that of Y-STRs with increased mutation rates, such as in RMplex.

Similarly, the differentiation rate for brother pairs were also improved strongly by both methods compared to those previously reported for sets with less Y-STRs. While Yfiler and the initial 13 RM Y-STRs previously differentiated 10.4% and 44.0% of brothers, respectively (2), based on the current data, Yfiler Plus and RMplex achieved a strong increase with differentiation rate estimates of 32.6% and 62.4%, respectively (Figure 1). As expected, the superiority of RMplex over Yfiler Plus observed for father-son pairs is also evident for brother pairs. Combining both methods increased the brother differentiation rate further to 72.9%. This increase was higher than that seen for father-son pairs based

on the combined marker set; however, in the current study, the sample size of brothers was with 92 pairs much smaller than that of father-son pairs. Therefore, the obtained brother differentiation rates are expected to be less reliable than those for father-son pairs, as illustrated by the larger error bars in Figure 1. Therefore, until many more brother pairs are analyzed in the future, the brother differentiation rates reported here shall be treated with care.

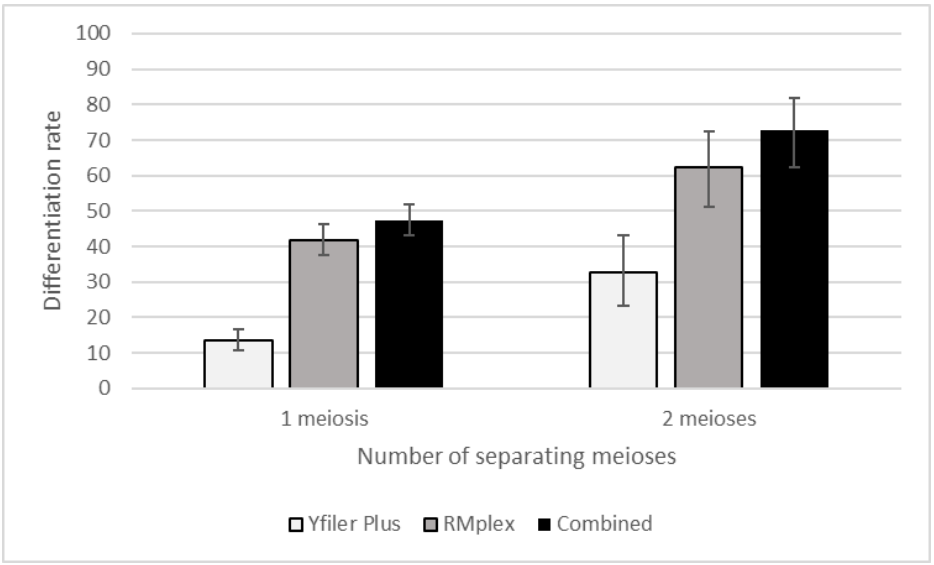


Figure 1: Male relative differentiation rates for father-son pairs (1 meiosis) and brother pairs (2 meioses) obtained with Yfiler Plus (25 Y-STRs), RMplex (30 Y-STRs), and both methods combined (49 Y-STRs). The error bars represent the exact binomial 95% confidence interval (Clopper-Pearson).

Since given the definition used, the above-described male relative differentiation rates are based on one or more mutations, we also investigated the number of Y-STRs that showed a mutational difference between any given differentiated father-son or brother pair (Figure 2, for illustrative reason we also included the pairs with zero mutations). Of all 209 father-son pairs differentiated with RMplex, 75.1% were separated by a single mutation at one Y-STR marker, 20.6% by mutations at two, and 4.3% at three markers. For the 71 father-son pairs differentiated with Yfiler Plus, the relative difference between the percentage of pairs differentiated by one mutation (94.4%), compared to those explained by mutations at two (4.2%) and three markers (1.4%), was larger than with RMplex, likely due to the lower number of Y-STRs with high mutation rates in Yfiler Plus. When considering all 49 Y-STRs targeted by both methods together, slightly more pairs were

differentiated by more than a single mutation (Figure 2). Notably, based on the combined analysis, six (2.5%) of the 237 differentiated father-son pairs showed mutations at four Y-STR markers, which was not seen with RMplex and Yfiler Plus alone. These data, demonstrating DNA-confirmed father-son pairs with mutations at 2-4 Y-STRs, provide relevant knowledge that shall be considered in interpreting results of Y-STR based paternity testing in deficiency cases where the mother of the disputed male child is unavailable for DNA testing and thus Y-STR testing is indicated.

Compared to father-son pairs, brother pairs separated by more than one mutation were seen more frequently for both methods separately and combined (Figure 2), which may be explained by the double number of meiosis separating brothers relative to father-sons and thus double the opportunity for a mutation to arise at any given locus. In contrast to father-son pairs, we did not observe a brother pair with more than three mutations, which could be a result of the smaller sample size of brothers relative to father-son pairs. Observing more than a single mutation in brothers was a lot more common with RMplex compared to Yfiler Plus in both father-son pairs and brother pairs (Figure 2).

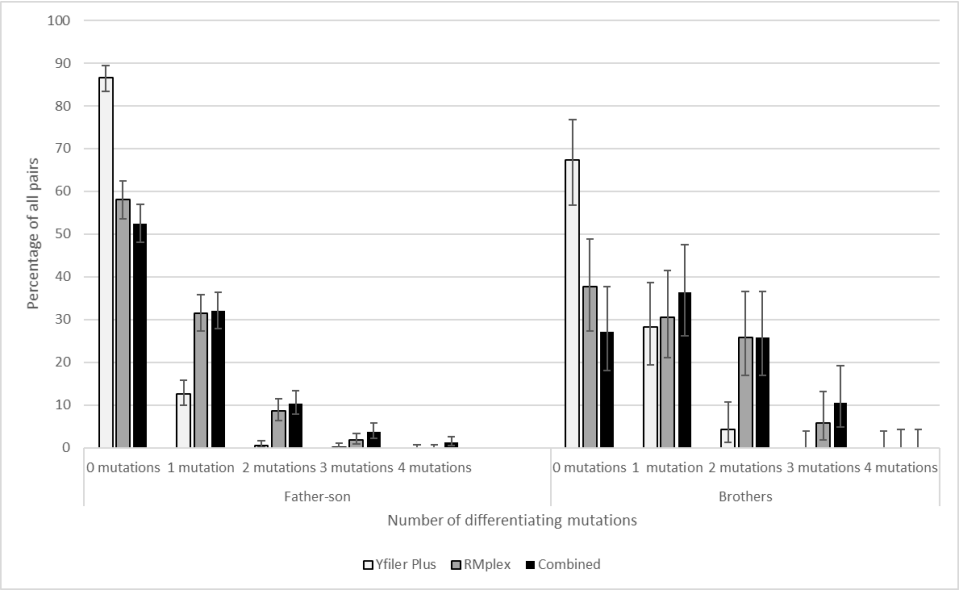


Figure 2. Proportions of father-son pairs and brother pairs analyzed with Yfiler Plus (25 Y-STRs), RMplex (30 Y-STRs), and both methods combined (49 Y-STRs) with mutations at zero, one, two, three, and four Y-STR markers per pair. None of these pairs was differentiated by mutations at more than four markers. The error bars represent the exact binomial 95% confidence interval (Clopper-Pearson).

Differentiating unrelated males

RM Y-STRs have previously been shown to not only increase the differentiation of related males, but also that of unrelated ones, e.g., when comparing results from the set of 13 initial RM Y-STRs with those from Yfiler targeting 16 Y-STRs in the same set of samples (4). Since in the current study, markers were added to both sets, we additionally investigated the differentiation of the unrelated individuals. For RMplex, we obtained 469 unique haplotypes among 470 unrelated men of which 468 were found in a single man and one was shared between two men, resulting in a haplotype discrimination capacity of 99.8%. For Yfiler Plus, we observed 481 unique haplotypes among 482 unrelated men of which 480 were found in a single man and one was shared between two men, which also results in a haplotype discrimination capacity of 99.8%. The estimated haplotype diversity for both marker sets was also the same with 0.999991. Notably, the two men that shared the same 30-marker RMplex haplotype also shared the same 25-marker YFiler Plus haplotype. Sharing the same allele at 49 Y-STRs, of which the majority has high mutation rates, is unlikely to be found in two unrelated men and likely indicates a hidden relationship. As a result of the complete sample anonymization prior to this study, it could not be investigated to what degree these two men are paternally related, but our Y-STR results strongly suggest that they are paternally related and rather closely than distantly related.

A previous study based on the 13 initial RM Y-STRs and the 16 Yfiler Y-STRs (4) reported an increase in haplotype diversity with RM Y-STRs relative to Yfiler for three population samples from Austria i.e., Tyrol, Upper Australia, and Salzburg (13 RM Y-STRs: 0.99988, 0.99996, and 0.99995, respectively; Yfiler: 0.9996, 0.9998, and 0.9998 respectively (4)). Also, in these previously analyzed Austrian population samples, the haplotype discrimination capacities were increased considerably with RM Y-STRs relative to Yfiler (13 RM Y-STRs: 99.2%, 99.6%, and 99.5%, Yfiler: 97.7%, 97.3% and 98.1%). The increased haplotype diversity and discrimination capacity the current (an albeit different) Austrian population sample from Salzburg reveals for both Y-STR sets is expected given that both sets contain more Y-STRs relative to their counterparts used in the previous study.

What was not expected, however, is that the present study did not reveal any differences in discrimination capacity and haplotype diversity between RMplex and Yfiler Plus. It may be that this equal finding is influenced by sample size effects, as the more Y-STRs are analyzed, and especially the more markers with high mutation rates, the larger the population sample size needs to be to obtain reliable diversity estimates. Until data

from more diverse populations and with larger sample size become available based on both methods, the present result of achieving the same differentiation of unrelated males with RMplex and Yfiler Plus shall be treated with care. In the future, it would be interesting to increase the sample size to see if the equal diversity measures obtained here for RMplex and Yfiler Plus remain or not. Our finding may also be influenced by the European population background of the samples analyzed here and future studies should investigate non-European populations. To allow future comparisons of the allele frequencies of these Y-STRs in different populations, we present the observed allele frequencies obtained from all fathers in the present study in supplementary Table S3.

Conclusions

We present here the first application of RMplex and Yfiler Plus on a relatively large set of DNA-confirmed father-son pairs for obtaining empirical estimates of mutation rates, male relative differentiation rates for father-sons and for brothers, as well as haplotype diversity based on the unrelated men. The mutation rates achieved here were not significantly different from those previously obtained for these markers and the established updated reference mutation rate estimates are made available for future use. Father-son and brother differentiation rates are reported here for the first time for both marker sets and methods. Also the first time, we empirically demonstrate the improved differentiation of close male relatives achieved with RMplex compared to the current state-of-the-art commercial Y-STR kit: Yfiler Plus. With the increased number of RM Y-STRs included in RMplex, our study reaffirms the high value of RM Y-STRs for differentiating paternally related males. Future work should perform RMplex analysis in more distantly related males and in male relatives with more diverse paternal biogeographical backgrounds. Motivated by our findings, we encourage the forensic Y-chromosome community to use RMplex in all forensic cases where a match with any previously or currently available commercial Y-STR kit was obtained between the male suspect and the evidence material, or to solely use RMplex in suitable cases such as sexual assault cases, aiming to find out if the male suspect left the evidence material at the crime scene, or any of his male paternal relatives did.

References

1. Sibille I, Duverneuil C, De La Grandmaison GL, Guerrouache K, Teissiere F, Durigon M, et al. Y-STR DNA amplification as biological evidence in sexually assaulted female victims with no cytological detection of spermatozoa. Forensic Science International. 2002;125(2-3):212-6.
2. Adnan A, Ralf A, Rakha A, Kousouri N, Kayser M. Improving empirical evidence on differentiating closely related men with RM Y-STRs: a comprehensive pedigree study from Pakistan. Forensic Science International: Genetics. 2016;25:45-51.
3. Adnan A, Rakha A, Lao O, Kayser M. Mutation analysis at 17 Y-STR loci (Yfiler) in father-son pairs of male pedigrees from Pakistan. Forensic Science International: Genetics. 2018;36:e17-e8.
4. Ballantyne KN, Ralf A, Aboukhalid R, Achakzai NM, Anjos MJ, Ayub Q, et al. Toward Male Individualization with Rapidly Mutating Y-Chromosomal Short Tandem Repeats. Human Mutation. 2014;35(8):1021-32.
5. Della Rocca C, Alladio E, Barni F, Cannone F, D'Atanasio E, Trombetta B, et al. LOW DISCRIMINATION POWER OF THE YFILER™ PLUS PCR AMPLIFICATION KIT IN AFRICAN POPULATIONS. DO WE NEED MORE RM Y-STRs? Forensic Science International: Genetics Supplement Series. 2019;7(1):671-3.
6. Purps J, Siegert S, Willuweit S, Nagy M, Alves C, Salazar R, et al. A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. Forensic Science International: Genetics. 2014;12:12-23.
7. Caliebe A, Jochens A, Willuweit S, Roewer L, Krawczak M. No shortcut solution to the problem of Y-STR match probability calculation. Forensic Science International: Genetics. 2015;15:69-75.
8. Andersen MM, Caliebe A, Jochens A, Willuweit S, Krawczak M. Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory. Forensic Science International: Genetics. 2013;7(2):264-71.
9. Andersen MM, Balding DJ. Assessing the forensic value of DNA evidence from Y chromosomes and mitogenomes. Genes. 2021;12(8):1209.
10. Andersen MM, Balding DJ. How convincing is a matching Y-chromosome profile? PLOS Genetics. 2017;13(11):e1007028.
11. Roewer L, Andersen MM, Ballantyne J, Butler JM, Caliebe A, Corach D, et al. DNA Commission of the International Society of Forensic Genetics (ISFG): Recommendations on the Interpretation of Y-STR results in Forensic Analysis. Forensic Science International: Genetics. 2020:102308.
12. Larmuseau MHD, Vanderheyden N, Van Geystelen A, Decorte R. A substantially lower frequency of uninformative matches between 23 versus 17 Y-STR haplotypes in north Western Europe. Forensic Science International: Genetics. 2014;11:214-9.
13. Coble MD, Hill CR, Butler JM. Haplotype data for 23 Y-chromosome markers in four US population groups. Forensic Science International: Genetics. 2013;7(3):e66-e8.
14. Ballantyne KN, Keerl V, Wollstein A, Choi Y, Zuniga SB, Ralf A, et al. A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. Forensic Science International: Genetics. 2012;6(2):208-18.
15. Ballantyne KN, Goedbloed M, Fang R, Schaap O, Lao O, Wollstein A, et al. Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. The American Journal of Human Genetics. 2010;87(3):341-53.
16. Ralf A, Lubach D, Kousouri N, Winkler C, Schulz I, Roewer L, et al. Identification and characterization of novel rapidly mutating Y-chromosomal short tandem repeat markers. Human Mutation. 2020;41(9):1680-96.

17. Ralf A, Zandstra D, Weiler N, van Ijcken WFJ, Sijen T, Kayser M. RMplex: An efficient method for analyzing 30 Y-STRs with high mutation rates. Forensic Science International: Genetics. 2021(55).
18. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 1934;26(4):404-13.
19. Fisher RA. Statistical methods for research workers. Breakthroughs in Statistics: Springer; 1992. p. 66-70.
20. Bonferroni C. Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze. 1936;8:3-62.
21. Gopinath S, Zhong C, Nguyen V, Ge J, Lagacé RE, Short ML, et al. Developmental validation of the Yfiler® Plus PCR Amplification Kit: An enhanced Y-STR multiplex for casework and database applications. Forensic Science International: Genetics. 2016;24:164-75.
22. Ambrosio IB, Braganholi DF, Orlando LBM, Andrekenas NC, da Mota Pontes I, da Silva DA, et al. Mutational data and population profiling of 23 Y-STRs in three Brazilian populations. Forensic Science International: Genetics. 2020;48:102348.
23. Bugoye FC, Mulima E, Misinzo G. Analysis of mutation rate of 17 Y-chromosome short tandem repeats loci using Tanzanian father-son paired samples. Genetics research international. 2018;2018.
24. Chen Y, Zhou W, Li M, Li Y, Huang L, Jiang L, et al. Mutation rates of 13 RM Y-STRs in a Han population from Shandong province, China. Forensic Science International: Genetics Supplement Series. 2017;6:e346-e8.
25. Da Fré NN, Rodenbusch R, Gastaldo AZ, Hanson E, Ballantyne J, Alho CS. Genetic data and de novo mutation rates in father-son pairs of 23 Y-STR loci in Southern Brazil population. International journal of legal medicine. 2015;129(6):1221-3.
26. Fan G, Pan L, Tang P, Zhou Y, Liu M, Luo X. developmental validation of a novel 41-plex Y-STR system for the direct amplification of reference samples. International Journal of Legal Medicine. 2021;135(2):409-19.
27. Fan H, Zeng Y, Wu W, Liu H, Xu Q, Du W, et al. The Y-STR landscape of coastal southeastern han: Forensic characteristics, haplotype analyses, mutation rates, and population genetics. Electrophoresis. 2021.
28. Fu J, Cheng J, Wei C, Khan MA, Jin Z, Fu J. Assessing 23 Y-STR loci mutation rates in Chinese Han father–son pairs from southwestern China. Molecular Biology Reports. 2020;47(10):7755-60.
29. Ge J, Budowle B, Aranda XG, Planz JV, Eisenberg AJ, Chakraborty R. Mutation rates at Y chromosome short tandem repeats in Texas populations. Forensic Science International: Genetics. 2009;3(3):179-84.
30. Hedman M, Neuvonen AM, Sajantila A, Palo JU. Dissecting the Finnish male uniformity: the value of additional Y-STR loci. Forensic Science International: Genetics. 2011;5(3):199-201.
31. Laouina A, Nadifi S, Boulouiz R, El Arji M, Talbi J, El Houate B, et al. Mutation rate at 17 Y-STR loci in "Father/Son" pairs from moroccan population. Legal Medicine. 2013;15(5):269-71.
32. Lin H, Ye Q, Tang P, Mo T, Yu X, Tang J. Analyzing genetic polymorphism and mutation of 44 Y-STRs in a Chinese Han population of Southern China. Legal Medicine. 2020;42:101643.
33. Liu J, Wang R, Shi J, Cheng X, Hao T, Guo J, et al. The construction and application of a new 17-plex Y-STR system using universal fluorescent PCR. International Journal of Legal Medicine. 2020;134(6):2015-27.
34. Oh YN, Lee HY, Lee EY, Kim EH, Yang WI, Shin K-J. Haplotype and mutation analysis for newly suggested Y-STRs in Korean father–son pairs. Forensic Science International: Genetics. 2015;15:64-8.
35. Otagiri T, Sato N, Shiozaki T, Harayama Y, Hayashi T, Kobayashi K, et al. Mutation analysis for 25 Y-STR markers in Japanese population. Legal Medicine. 2021;50:101860.

36. Petrovic V, Kecmanovic M, Markovic MK, Keckarevic D. Assessment of mutation rates for PPY23 Y chromosome STR loci in Serbian father-son pairs. Forensic Science International: Genetics. 2019;39:e5-e9.
37. Rogalla U, Woźniak M, Swobodziński J, Derenko M, Malyarchuk BA, Dambueva I, et al. A novel multiplex assay amplifying 13 Y-STRs characterized by rapid and moderate mutation rate. Forensic Science International: Genetics. 2015;15:49-55.
38. Sánchez ME, Burgos G, Gaviria A, Aguirre V, Vela M, Leone PE, et al. Y STRs mutation events in father-son pairs in Ecuadorian individuals. Forensic Science International: Genetics Supplement Series. 2015;5:e310-e1.
39. Serin A, Ay M, Sevay H, Gurkan C, Canan H. Genetic characterisation of 13 rapidly mutating Y-STR loci in 100 father and son pairs from South and East Turkey. Annals of Human Biology. 2018;45(6-8):506-15.
40. Wang Y, Zhang Y-j, Zhang C-c, Li R, Yang Y, Ou X-l, et al. Genetic polymorphisms and mutation rates of 27 Y-chromosomal STRs in a Han population from Guangdong Province, Southern China. Forensic Science International: Genetics. 2016;21:5-9.
41. Wang Q, Jin B, An G, Zhong Q, Chen M, Luo X, et al. Rapidly mutating Y-STRs study in Chinese Yi population. International Journal of Legal Medicine. 2019;133(1):45-50.
42. Weng W, Liu H, Li S, Ge J, Wang H, Liu C. Mutation rates at 16 Y-chromosome STRs in the South China Han population. International journal of legal medicine. 2013;127(2):369-72.
43. Wu W, Ren W, Hao H, Nan H, He X, Liu Q, et al. Mutation rates at 42 Y chromosomal short tandem repeats in Chinese Han population in Eastern China. International journal of legal medicine. 2018;132(5):1317-9.
44. Yang Y, Wang W, Cheng F, Chen M, Chen T, Zhao J, 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):1-6.
45. Yuan L, Chen W, Zhao D, Li Y, Hao S, Liu Y, et al. Mutation analysis of 13 RM Y-STR loci in Han population from Beijing of China. International Journal of Legal Medicine. 2019;133(1):59-63.
46. Zhang W, Xiao C, Yu J, Wei T, Liao F, Wei W, et al. Multiplex assay development and mutation rate analysis for 13 RM Y-STRs in Chinese Han population. International Journal of Legal Medicine. 2017;131(2):345-50.
47. Zhou Y, Song F, Dai H, Wang S, Zhang K, Wei X, et al. Developmental validation of the Microreader™ RM-Y ID System: a new rapidly mutating Y-STR 17-plex system for forensic application. International Journal of Legal Medicine. 2021:1-12.
48. Claerhout S, Vandenbosch M, Nivelle K, Gruyters L, Peeters A, Larmuseau MHD, et al. Determining Y-STR mutation rates in deep-routing genealogies: Identification of haplogroup differences. Forensic Science International: Genetics. 2018;34:1-10.

Supplementary information

Out of environmental considerations the supplementary materials belonging to this publication were not printed with this chapter of the thesis. The digital files can be obtained with the original publication at:
https://doi.org/10.1016/j.fsigen.2022.102682

Chapter 5

RMplex reveals population differences in RM Y-STR mutation rates and provides improved father-son differentiation in Japanese

Tomomi Otagiri[1], Noriko Sato[1], Hideki Asamura[1], Evelina Parvanova[2], Manfred Kayser[2], Arwin Ralf[2]

[1] Department of Legal Medicine, Shinshu University School of Medicine, Matsumoto, Nagano, Japan
[2] Department of Genetic Identification, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands

Published in revised form in *Forensic Science International: Genetics*, August 2022
Volume 61, 102766
doi: https://doi.org/10.1016/j.fsigen.2022.102766

Abstract

Rapidly mutating Y-chromosomal short tandem repeat markers (RM Y-STRs) characterized by increased mutation rates are suitable for differentiating both related and unrelated males, as is relevant in forensic casework where autosomal STR profiling is unavailable and only Y-STR analysis can be done successfully. The recently introduced multiplex method RMplex allows for the efficient analysis of 30 Y-STRs with increased mutation rates, including all 26 currently known RM Y-STRs. While currently available RM Y-STR mutation rates and male relative differentiation rates were established mostly from European data, here we applied RMplex to DNA samples of genetically confirmed father-son pairs from East Asia. For several RM Y-STRs, we found significantly higher mutation rates in Japanese compared to previous estimates. The consequent father-son differentiation rate based on RMplex was significantly higher (52%) in Japanese than previously reported for Europeans (42%), and much higher than with Yfiler Plus in both sample sets (14% and 13%, respectively). We demonstrate that the higher mutation and relative differentiation rates in Japanese is likely explained by on average longer Y-STR alleles relative to Europeans. Moreover, we show that the most striking differences, which were found in DYS712, could be linked to a Y-SNP haplogroup (O1b2-P49) that is common in Japanese and rare in other populations. Our results are in line with the hypothesis of population founder and isolation effects as underlying reason for differences in mutation rates between populations, which we confirm here and additionally also demonstrate for the relative differentiation rate. We encourage the forensic Y-STR community to generate more RMplex data from more population samples of sufficiently large samples size in combination with Y-SNP data to further investigate population effects on mutation and relative differentiation rates. Until more RMplex data from more populations become available, caution shall be placed when applying RM Y-STR mutation rate estimates established in one population, such as Europeans, to forensic casework involving male suspects of paternal origin from other populations, such as non-Europeans.

Introduction

Y chromosomal short tandem repeats (Y-STRs) are commonly used in forensic DNA analysis, especially in sexual assault cases involving mixed DNA evidence to which the male perpetrator and the female victim contributed, as such male-female mixed DNA samples are notorious for showing difficulties in perpetrator identification based on autosomal STR profiling [1]. However, due to the paternal inheritance of the male-specific part of the Y-chromosome, Y-STR profiles are typically shared between paternally related men. In consequence, a Y-STR haplotype match between the male suspect and the evidence DNA does not necessarily indicate that the suspect contributed to the evidence sample, as this could also have been any of his male relatives sharing the same Y-STR haplotype. The need to overcome this limitation of differentiating paternally related males with Y-STRs motivated the search for, and led to the identification of, a new group of Y-STRs characterized by elevated mutation rates [2], which had mutation rates of 10^{-2} (1 mutation every 100 generations per locus, mpg) and higher, and were termed rapidly mutating Y-STRs (RM Y-STRs) [3]. Recently, a new set of RM Y-STRs was discovered by applying in silico search for candidate markers and empirical confirmation in father-son pairs [4], which increased the number of currently known RM-Y-STRs to 26 and also identified Y-STRs with mutation rates higher than 10^{-3} but lower than 10^{-2} termed fast mutating Y-STRs (FM Y-STRs). Moreover, a new multiplex genotyping assay, RMplex, was developed and validated that analyses a total of 30 Y-STRs with increased mutation rates, which includes the 26 RM Y-STRs and 4 FM-Y-STRs [5]. It was shown that this new RMplex assay outperformed the state-of-the art commercial Y-STR genotyping assay Yfiler™ Plus PCR Amplification Kit, demonstrating increased father-son differentiation rate by a factor of three [6], underlining the suitability of both, RM Y-STRs as markers and RMplex as methodology for differentiating paternally related men as relevant in forensic casework.

Notably, all currently known RM Y-STRs were discovered solely via mutation rate studies in Europeans [2, 4]. While some worldwide father-son and population data were previously established for the initial set of 13 RM Y-STRs [7], non-European data for the new set of 26 RM Y-STRs and for the full set of 30 Y-STRs included in RMplex are missing as of yet. Some previous studies showed differences in mutation rates between males from populations for the initial set of 13 RM Y-STRs [7-13]. While most previous Y-STR mutation rate studies lack Y-SNP haplogroup data, Claerhout et al. studied the effect of Y-SNP haplogroup related Y-STR allele frequency distributions on Y-STR mutation rates for European males [14] and found lower overall average mutation rates within some Y-SNP haplogroups (i.e., haplogroup I & J) compared to another others (i.e., R1b). Lower Y-STR mutation rates coincided with allele frequency distributions that were skewed towards

shorter alleles for some Y-STRs in males with haplogroup I & J. On the other hand, previous Y-STR mutation studies revealed a strong impact of allele length, i.e. number of repeats, on mutability of Y-STRs, with longer alleles leading to increased and shorter ones to decreased mutation rates [15, 16]. Many populations have varying Y-SNP haplogroup compositions indicating different male founders who carried their own Y-STR haplotypes. Provided strong enough founder and genetic isolation effects, this may explain the present-day Y-STR allelic distribution differences seen between populations characterized by strong Y-SNP haplogroup differences. Given the well-known impact of repeat length on Y-STR mutability [15, 16], male founders with longer Y-STR alleles may lead to increased mutability of such Y-STRs in the population of descendants, while founders with shorter Y-STR alleles may lead to decreased mutability in such a population, but data evidence to empirically prove this hypothesis is scarce.

Other factors than population effects may also be at the basis of differences in observed mutation rates between different populations. For instance, Y-STR mutations are rare events and thus mutation rate estimates are prone to stochastic effects. Hence, ideally, thousands of father-son pairs should be analyzed to make accurate Y-STR mutation rate estimates in samples from many populations; in reality, however, most studies only involve hundreds and only a limited number of populations had been analyzed thus far. Based on the underlying mutation rates, analyzing some hundreds of pairs, results in only a small number of mutations per Y-STR locus observed and thus unreliable mutation rate estimates. Limited sample sizes can therefore easily lead to what seems to be large differences in Y-STR mutation rates, but actually reflect chance effects. Obviously, the larger the mutation rate of a Y-STR, the more mutations will be observed, thereby decreasing the impact of stochastic effects, also in studies of limited sample size. This makes RM Y-STRs the most suitable markers, and the large number of Y-STRs with increased mutation rates included in RMplex makes RMplex the most suitable methodology for studying population effects on Y-STR mutation and male relative differentiation rates.

A recent study applied the current state-of-the-art commercial Yfiler™ Plus PCR Amplification Kit (Yfiler) to Japanese father-son pairs and found that the six RM Y-STR markers included in this commercial kit did not show the expected elevated mutation rates known from previous European studies [17]. The authors therefore concluded that this commercial kit may be less suitable for the purpose of male relative differentiation in the Japanese population. While many other explanations may allow interpreting these findings in alternative ways, population effects on Y-STR mutation rates particularly seen at RM Y-STRs could be one of the reasons for these findings, but with only 6 RM Y-STRs analyzed such conclusion was difficult to derive from this previous study.

In the current study, we applied RMplex to DNA-confirmed father-son pairs originating from Japan. The newly established mutation and relative differentiation rate data were compared to previously published consensus mutation rate estimates based on multiple studies [6]. Additionally, Y haplogroup data were generated from genotyping 10 Y-SNPs to infer 8 Y haplogroups known to be frequent in Japanese, which allowed us to link the established Y-STR mutation data with the relevant Y-SNP haplogroup background information.

Materials and methods

DNA samples

Genomic DNA was extracted of a total of 340 males using QIAamp DNA Blood Mini Kit (Qiagen) following the manufacturer's instructions. Of the 340 males were included in this study, 296 originated from pairs of fathers with a single son (n=148), while additionally the dataset consisted of one father with three sons, and 12 fathers with two sons. Lastly, the dataset included one family with four generations of male relatives (one individual per generation, three father-son pairs). In total, 178 father-son pairs were defined from this dataset. The biological relationship of all pair of males was confirmed through autosomal STR typing. This study was approved by the Ethics committee of Shinshu University School of Medicine (Permission number: 667). All individuals included in this study provided informed consent. A total of 280 samples used in the present study overlap with those in the previous study based on Yfiler Plus [17], while the remaining 60 samples were newly typed in the context of the present study.

Amplification and genotyping

Y-STR typing was done using the protocol previously described by Ralf et al. [5], while using the alternative forward primer for DYS570 as described in that publication. PCR was performed in 10 µL reaction volumes. Amplification was performed using a GeneAmp® PCR System 9700 (Thermo Fisher Scientific), the resulting PCR products were analyzed using a 3500 Genetic Analyzer (Thermo Fisher Scientific) and the resulting electropherograms were analyzed using GeneMapper® ID-X Software v1.4 (Thermo Fisher Scientific).

Y-SNP based haplogroup analysis

Y-SNP testing was performed by developing a custom-made multiplex genotyping assay using SNaPshot™ Multiplex Kit (Thermo Fisher Scientific). A total of 10 Y-SNPs were selected, 8 target specific (sub)haplogroups that were expected to be present in the Japanese male population: C-M130, D-M174, N-M231, O1a-M119, O1b2-P49, O2-M122, O-P186*(xM119,P49,M122), Q-M242). Additional two intermediate Y-SNPs were included in the design: DE-M145 and CF-P143. The primer sequences and thermal cycler conditions are shown in supplementary Table S1. The PCR was performed in 10 µL volumes including 5 µL QIAGEN Multiplex PCR Plus Kit (Qiagen), PCR at primers with the concentrations as detailed in supplementary Table S1 and 1 µL of DNA (~ 1 ng/ µL). Amplification was performed using a GeneAmp® PCR System 9700 (Thermo Fisher Scientific), the resulting PCR products were analyzed using a 3500 Genetic Analyzer (Thermo Fisher Scientific) and the resulting electropherograms were analyzed using GeneMapper® ID-X Software v1.4 (Thermo Fisher Scientific).

Data analysis

Both the mutation rates and the differentiation rates were calculated using the frequentist approach. The Clopper-Pearson interval was used to determine the 95% confidence intervals of these rates. Fisher's exact tests were used to determine the statistical significance of difference observed between the present study and previous studies. Pairwise Rst values were calculated by an in-house pipeline that performs per-marker allele comparisons for each pair of samples. The Rst value for a given pair was defined as the sum of the differences among all Y-STRs. Multi-copy markers pose additional complexity for this approach because it is typically not possible to tell which copies correspond to each other. To calculate the Rst value in such Y-STRs the pipeline chose the shortest path, e.g. if one individual typed with alleles 12, 16 was paired with another individual displaying alleles 13, 15, the pipeline derived a distance value of 2 (the sum of the difference between 12 and 13, and the difference between 15 and 16) instead of 6 (the sum of the difference between 12 and 15 and the difference between 13 and 16). R [18] was used to create boxplots.

Results and discussion

Mutation analysis

In total, 157 mutations were observed amongst the 178 Japanese father-son pairs, of which 138 were detected using RMplex and 29 with Yfiler Plus of which 10 were found at the six RM Y-STRs overlapping between the two methods. All mutations with the specific allele changes and the haplogroup of the pairs in which they occurred are shown in supplementary Table S2. Generally speaking, the sample size in the present study was, with 178 father-son pairs, relatively small to yield highly reliable estimations of the mutation rates. The uncertainty becomes apparent from the 95% confidence intervals we generated where even Y-STRs not showing a single mutation among the 178 pairs have an upper limit of 2 x 10^{-2} mpg (Table 1). Because of the limited sample size, all conclusions ought to be treated with caution, at least until larger scale studies in the same population replicate these results. The Y-STR marker that showed most mutations was the RM Y-STR DYF399S1 with 19 mutations alone, which confirms previous studies in different populations showing that this multi-copy RM Y-STR is the most mutable Y-STR [2, 4, 6-13]. However, in the current study DYF399S1 mutated even more frequently than previously observed with >10% of the analyzed pairs showing a mutation at this RM Y-STR (Table 1). The noted increase in mutation rate was statistically significant (p-value: 0.028) relative to the reference mutation rate based on >7,500 father-son pairs of which only 6.3% displayed a mutation for this multi-copy marker [6].

Another remarkable result was the high number of 17 mutations found at DYS712, which resulted in a mutation rate estimate of 9.6 x 10^{-2} mpg. This mutation rate we obtained for DYS712 here from Japanese father-son pair data is significantly higher than those previously obtained from European father-son pair data with 2.7 x 10^{-2} mpg (p-value: <0.0001) [4] and 4.3 x 10^{-2} mpg (p-value: 0.0138) [6]. A study analyzing father-son pairs from the Shanxi Province in China reported for DYS712 a mutation rate of 3.0 x 10^{-2} mpg [19], which is similar to the rates previously obtained from European data for this marker, but significantly lower than the rate we obtained here for Japanese (p-value: 0.0030). Although we cannot exclude the possibility that the increased mutation rate that was found for DYS712 in the present Japanese study relative to the previous European and Chinese studies represents a stochastic effect, resulting from the small sample size in the present study, the possibility that this Y-STR genuinely is more mutable in the Japanese population needs to be considered too. Following the population founder hypothesis, an explanation may be found in the distribution of the allele lengths. A more detailed reflection of this potential explanation will follow in section 3.4.

Notably, the commonly used Y-STR DYS458 showed a mutation rate of 3.4×10^{-2} mpg in the present study, which is remarkably higher compared to the rate of 8.0×10^{-3} mpg that was previously estimated in a study using European father-son pairs [2]. Also, compared to the consensus estimate based on data from 11,830 father son-pairs [6], the observed mutation rate in the current Japanese study is significantly higher (p-value: 0.0051) than the consensus mutation rate of 8.5×10^{-3} that was previously obtained (Table 1). A previous study of 213 Japanese father-son pairs that partly overlapped with the samples analyzed here only found a mutation rate of $1.4\ 10^{-2}$ [17], which underlines the stochastic effects that can occur when estimating mutation rates from relatively small numbers of father-son pairs.

Other Y-STRs that show significant difference compared to the previously reported consensus mutation rate estimates [6] are: DYF1001 (p-value 0.0120), DYS460 (p-value 0.0092), and DYS713 (p-value 0.0162). Notably, all these Y-STRs showed an increased mutation rate in the present study compared to the previous consensus estimates. Table 1 shows mutation rate estimates as obtained in the present study; these mutation rates are compared to the consensus Y-STR mutation rate estimates as described in Neuhuber et al., 2022 [6].

Table 1: Empirically established locus-specific mutation rates of 49 Y-STRs by applying RMplex and Yfiler™ Plus to a total of 178 DNA-confirmed father-son pairs from Japan and their comparisons with consensus locus-specific reference mutation rates previously described [6].

Marker	Assay	Total pairs	Mutations	Expansions	Contractions	Mutation rate ($x10^{-3}$)	95% confidence interval ($x10^{-3}$)	Reference mutation rate ($x10^{-3}$)	p-value[#]
DYF399S1	RMplex	178	19	7	12	106.7	65.5-161.7	62.8	**0.0280**
DYS712	RMplex	178	17	8	9	95.5	56.6-148.5	31.1	**0.0001**
DYF1001	RMplex	178	17	11	6	95.5	56.6-148.5	48.0	**0.0120**
DYF403S1a	RMplex	178	9	5	4	50.6	23.4-93.8	27.3	0.0985
DYF1000	RMplex	178	8	4	4	44.9	19.6-86.6	35.9	0.5305
DYS713	RMplex	178	7	1	6	39.3	16-79.3	13.9	**0.0162**
DYS458	Yfiler Plus	178	6	3	3	33.7	12.5-71.9	8.5	**0.0051**
DYS724	RMplex	178	6	2	4	33.7	12.5-71.9	48.0	0.4643
DYS1010	RMplex	178	5	1	4	28.1	9.2-64.3	14.0	0.1852
DYS711	RMplex	178	4	2	2	22.5	6.2-56.5	26.6	1.0000
DYS612	RMplex	178	4	1	3	22.5	6.2-56.5	16.3	0.5401
DYF403S1b	RMplex	178	4	4	0	22.5	6.2-56.5	9.1	0.0859
DYS1005	RMplex	178	4	0	4	22.5	6.2-56.5	9.8	0.1191
DYS576	Yfiler Plus+RMplex	178	4	1	3	22.5	6.2-56.5	12.7	0.2938
DYS460	Yfiler Plus	178	4	2	2	22.5	6.2-56.5	4.3	**0.0092**
DYS547	RMplex	178	3	3	0	16.9	3.5-48.5	14.7	0.7474
DYS1007	RMplex	178	3	2	1	16.9	3.5-48.5	17.2	1.0000
DYF404S1	RMplex	178	3	1	2	16.9	3.5-48.5	12.5	0.4921
DYS1013	RMplex	178	2	1	1	11.2	1.4-40	10.8	1.0000
DYR88	RMplex	178	2	1	1	11.2	1.4-40.0	26.3	0.3170
DYS526b	RMplex	178	2	1	1	11.2	1.4-40.0	12.3	1.0000
DYF1002	RMplex	178	2	1	1	11.2	1.4-40.0	16.8	0.7650
DYS570	Yfiler Plus+RMplex	178	2	2	0	11.2	1.4-40.0	8.3	0.6612
DYS1003	RMplex	178	2	1	1	11.2	1.4-40.0	12.6	1.0000
DYS481	Yfiler Plus	178	2	1	1	11.2	1.4-40.0	4.7	0.2141
DYS1012	RMplex	178	2	1	1	11.2	1.4-40.0	15.8	1.0000
DYS626	RMplex	178	2	0	2	11.2	1.4-40.0	8.6	0.6676
DYS627	Yfiler Plus+RMplex	178	2	2	0	11.2	1.4-40.0	14.5	1.0000
DYS449	Yfiler Plus+RMplex	178	1	0	1	5.6	0.1-30.9	11.2	0.7258
DYS385	Yfiler Plus	178	1	1	0	5.6	0.1-30.9	7.5	1.0000
DYF387S1	Yfiler Plus+RMplex	178	1	0	1	5.6	0.1-30.9	10.2	1.0000
DYS456	Yfiler Plus	178	1	1	0	5.6	0.1-30.9	4.4	0.5443
DYS393	Yfiler Plus	178	1	1	0	5.6	0.1-30.9	1.7	0.2719
DYS19	Yfiler Plus	178	1	0	1	5.6	0.1-30.9	2.0	0.3041
YGATAH4	Yfiler Plus	178	1	1	0	5.6	0.1-30.9	1.9	0.2980
DYS442	RMplex	178	1	0	1	5.6	0.1-30.9	7.4	1.0000
DYS635	Yfiler Plus	178	1	1	0	5.6	0.1-30.9	3.8	0.5005
DYS389II	Yfiler Plus	178	1	0	1	5.6	0.1-30.9	5.5	0.6267
DYS533	Yfiler Plus	178	0	0	0	0.0	0-20.5	3.5	1.0000
DYS439	Yfiler Plus	178	0	0	0	0.0	0-20.5	4.8	1.0000
DYS391	Yfiler Plus	178	0	0	0	0.0	0-20.5	2.5	1.0000
DYS518	Yfiler Plus+RMplex	178	0	0	0	0.0	0-20.5	13.3	0.1784
DYS437	Yfiler Plus	178	0	0	0	0.0	0-20.5	1.2	1.0000
DYF393S1	RMplex	178	0	0	0	0.0	0-20.5	7.1	0.6248
DYS389I	Yfiler Plus	178	0	0	0	0.0	0-20.5	2.4	1.0000
DYS448	Yfiler Plus	178	0	0	0	0.0	0-20.5	0.8	1.0000
DYS390	Yfiler Plus	178	0	0	0	0.0	0-20.5	2.7	1.0000
DYS438	Yfiler Plus	178	0	0	0	0.0	0-20.5	0.3	1.0000
DYS392	Yfiler Plus	178	0	0	0	0.0	0-20.5	0.8	1.0000

statistically significant differences (p-values <0.05) are indicated in bold

Differentiation of father-son pairs

Based on Yfiler Plus alone, a total of 25 out of the 178 (14%) Japanese father-son pairs were differentiated, which is comparable to a previous study based on European males where 13% of father-son pairs were differentiated [6]. Based on RMplex alone, a total of 93 out of the 178 pairs (52%) were differentiated, which reflects an about 3.7-fold increase compared to Yfiler Plus in the same samples, and is significantly higher (Fisher's exact p-value: 0.0179) than in a previous study based on European father-son pairs where a differentiation rate of 42% was reported [6]. Combining the data from both Yfiler Plus and RMplex resulted in an even higher father-son differentiation rate of 57% with 101 out of the 178 Japanese pairs being separated. With less than a five percent point increase, the contribution of the non-overlapping Yfiler Plus markers to the differentiation of male relatives was limited, as expected based on their lower mutation rates.

When looking more closely to the number of mutations that differentiate a father-son pair, we saw that 22 of the 25 pairs differentiated with Yfiler Plus (88%) only showed a mutation at a single Y-STR marker. In contrast, for RMplex and for both assays combined, 62% of the 93 and 57% of the 101 differentiated pairs were separated only by a single mutation, respectively. Furthermore, 8%, 27%, and 33% of the differentiated pairs were separated by mutations at two Y-STRs and 4%, 11%, and 8% of the differentiated pairs by mutations at three Y-STRs, for Yfiler Plus, RMplex, and the combined methods, respectively (Figure 1). Mutations at four and five markers were only observed when both methods were combined and were observed only in a single pair, respectively.

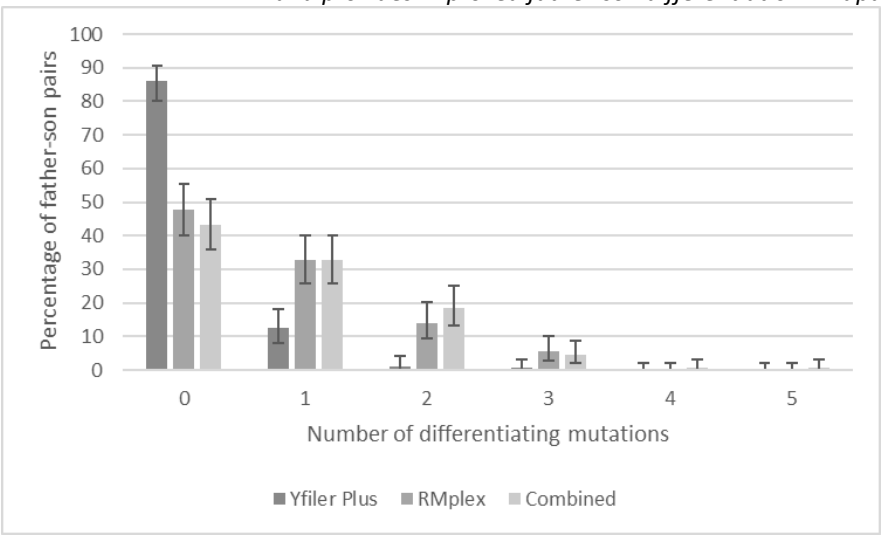


Figure 1. Percentage of father-son pairs analyzed with Yfiler Plus (25 Y-STRs), RMplex (30 Y-STRs), and both methods combined (49 Y-STRs) with mutations at zero, one, two, three, four, and five Y-STR markers per pair. None of these pairs was differentiated by mutations at more than five Y-STRs. The error bars represent the exact binomial 95% confidence interval (Clopper-Pearson).

Differentiation of unrelated males

To assess the efficiency in differentiated unrelated males, we compared the Y-STR haplotypes obtained with Yfiler Plus and RMplex combined in the total of 162 fathers and found that each of them carried a unique haplotype based on the full set of 49 Y-STR. Notably, the same number of unique haplotypes were also seen when considering RMplex and Yfiler Plus separately. Hence, in the current study, no difference in capabilities to differentiate unrelated males was seen for RMplex and Yfiler Plus, which was also reported in a previous European study [6]. However, the relatively low sample size in both studies might have influenced this rather unexpected result, as the probability of observing shared haplotypes too increases with sample size. Future studies with increased sample size need to show if indeed both methods are performing equally well in differentiating unrelated men, or if the identical performance of both methods was influenced by sample size effects in this Japanese and the previous European study.

However, despite its limited sample size, the data obtain in the current study does provide some insights in the potential to differentiate unrelated males by determining pair-wise Rst values. Rst considers the mutational differences between haplotypes and is typically estimated from data of different population to express the proportion of the diversity seen between populations. Here we estimated Rst between

pairs of individual haplotypes, and not between populations as typically done, and this way, Rst provides an estimate on the diversity difference between the haplotypes derived from the two genotyping methods. As shown in Figure 2, there is a sharp distinction between Yfiler Plus and RMplex in pairwise-Rst distributions, where Yfiler Plus clearly results in lower Rst values, indicating more similarity between the Yfiler Plus derived haplotypes from the unrelated males compared to those from RMplex. By extrapolation, it could be expected that more similarity in a small sample would translate to more overlapping haplotypes in a significantly lager sample. Although, there is still a need for empirical evidence based on large numbers of unrelated males, the difference in Rst values between the two methods could be seen as a first indication that RMplex may be superior in differentiating not only related males but also unrelated ones.

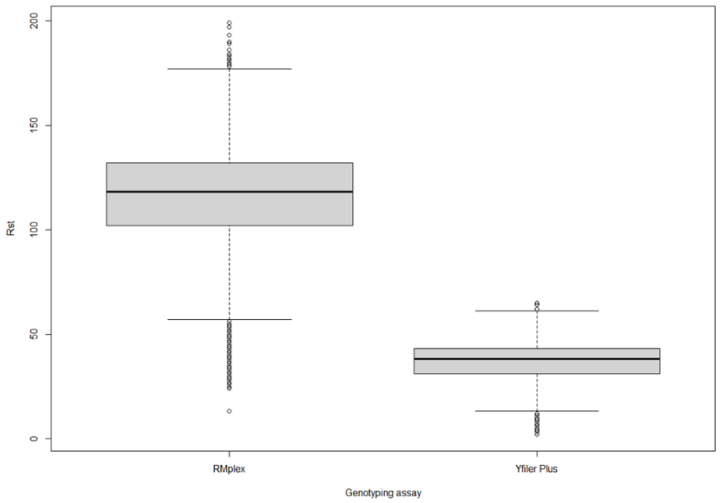


Figure 2. Pairwise Rst-value distribution obtained RMplex and Yfiler Plus based on the 162 unrelated Japanese males.

Differences in Y-STR allele lengths between populations and Y-SNP haplogroups

Y-SNP analysis was performed on 162 unrelated fathers with one individual not providing a full profile, leaving 161 unrelated males in this analysis for which Y-SNP based haplogroups were established. Selection of the 10 Y-SNPs was done to allow detecting

most common haplogroups known for Japanese. The most commonly observed haplogroup in this Japanese dataset was O1b2-P49 (32%), followed by D-M174 (30%), O2-M122 (19%), C-M130 (14%). The remaining four haplogroups were found in less than 5 individuals each (<2.5%). A previous study that analyzed Y-SNPs in commonly observed Japanese surnames found that 37% belonged to haplogroup D, 30% to O1b, 20% to O2, and 9% to C [20]. Despite some stochastic variations, these results show a similar occurrences of the most common haplogroups in the Japanese population.

One of the most remarkable results obtained in the present study was the increased number of mutations found at DYS712. As hypothesized before, this may be the result of high prevalence of longer alleles for this Y-STR in the Japanese population compared to, for example, European populations. To test this hypothesis, we compared the allele frequencies previously reported for DYS712 from a European population sample from Austria [6], to the allele frequencies found in the current Japanese samples. Figure 3 shows that there was more variability in the DYS712 Y-STR alleles found in the Japanese relative to the Europeans; moreover, longer alleles were more common in the Japanese samples relative to the Europeans, and the Japanese sample had a higher median allele length compared to the European. When comparing the most common Y-SNP haplogroups in the Japanese sample set, it becomes evident that the longer Y-STR alleles at DYS712 were especially common in males belonging to a subgroup of haplogroup O1b2 (O-P49), which was the most frequently observed haplogroup in the current Japanese sample and is completely absent from Europeans [21]. It is widely established that longer Y-STR alleles are more prone to mutations [15, 16]. This notion is further supported by the fact that in the current study 13 out of the total 17 mutations (76%) observed at DYS712 arose from fathers that had an allele length larger than the population median value of 23 in this Japanese dataset. That this effect likely is haplogroup dependent is further supported by our finding that 9 out of 17 pairs (53%) that showed a mutation for DYS712 belonged to haplogroup O1b2 (O-P49). Notably, this haplogroup was rarely found in Eastern Han Chinese [22], which may explain why the high mutation rate found in the current Japanese dataset for DYS712 was not previously found in Chinese from the Shanxi region [23]. Also the opposite effect is seen in the Japanese data, where the frequently occurring haplogroup D generally displays short Y-STR alleles at DYS712 (Figure 2), while only 1 out of the 17 pairs (6%) that showed a mutation at DYS712 belonged to haplogroup D despite the high overall prevalence of haplogroup D in the studied samples. All allele frequencies for all 49 Y-STRs analyzed with RMplex and Yfiler Plus are shown in supplementary Table S3.

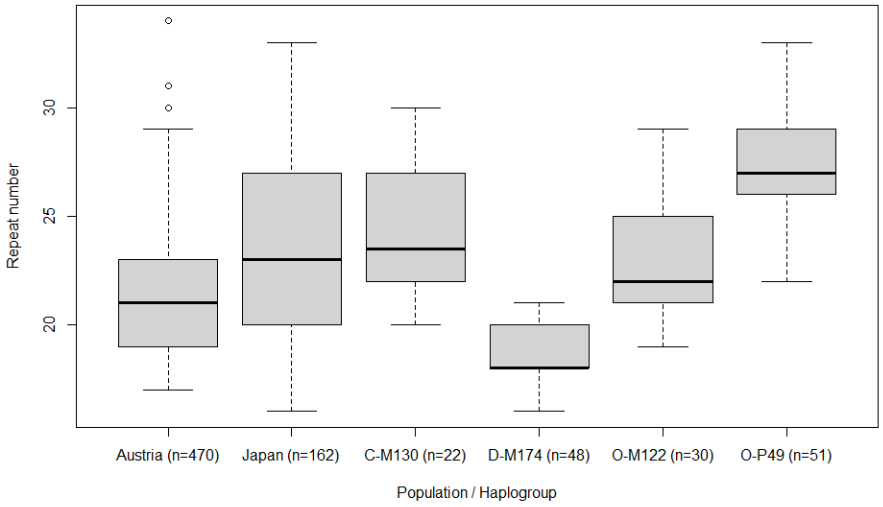


Figure 3. Boxplots showing the difference in allele length distribution of the RM Y-STR marker DYS712 between the current Japanese study (overall and stratified per predicted Y-SNP haplogroup) where this marker has a mutation rate of 9.6×10^{-2} and a previous European study where the mutation rate was 3.1×10^{-2} [6].

To further investigate the potential influence of allele frequencies on mutation rates, we compared both the mutation rates and the mean allele frequency from both the current Japanese study and the previous European study [6] for the six Y-STRs that showed significantly different mutation rates between the current study and the previously established consensus mutation rates (Table 1). As evident from Table 2, the observed mutation rate differences were not always statistically significant. Both studies had a relatively small sample size, whereas the sample size of the consensus mutation rate estimates used in prior comparisons were relatively large. Regardless, the higher occurrence of mutations in the current study on Japanese father-son pairs compared to the study based on European pairs was still clearly noticeable (Table 2). Interestingly, for all six Y-STRs, the mean allele frequency was also higher in the Japanese population than it was in the European population, albeit differently so across the six markers (Table 2).

Table 2: Direct comparison of the mutation rates and mean allele frequencies between the current Japanese study and a previous European study [6] for six Y-STRs with remarkable high mutation rate estimates in the current Japanese study.

	Mutation rates (x10^{-3})			Mean allele frequency	
Marker	**Japanese**	**European**	**p-value#**	**Japanese**	**European**
DYF399S1	106.7	77.4	0.217	23.9	23.1
DYF1001	95.5	36.0	0.005	64.3	62.5
DYS712	95.5	43.4	0.014	23.4	21.4
DYS713	39.3	17.0	0.139	46.1	43.0
DYS458	33.7	13.2	0.103	17.0	16.2
DYS460	22.5	7.5	0.115	10.7	10.5

The p-values are based on Fisher's exact tests and the proportions of the father-son pairs where mutations were observed from each study, respectively.

Due to limited sample size, the evidence we present here for increased RM Y-STR mutation and relative differentiation rates in Japanese relative to Europeans being explained by population effects is statistically not clear-cut. However, our findings show a trend that supports the population founder hypothesis to explain Y-STR mutation rate differences between populations, which shall be confirmed with more Japanese data in future studies. Moreover, also other population samples of paternally related males from different parts of the world and of suitable size should be analyzed with RMplex and Y-SNP haplogroups to collect more empirical evidence on whether population differences in RM Y-STR mutation and relative differentiation rates truly exist and can be explained.

Conclusion

Here, we showed for the first time, the efficiency at which RMplex differentiates non-European, i.e., Japanese, father-son pairs, which turned out to be at a significantly higher rate (52%) than previously established in Europeans (42%), and much higher than with the current state-of-the-art commercial Y-STR kit Yfiler Plus (14%). We also present significant differences in Y-STR mutation rates between Japanese and other populations, indicating

population effects influencing Y-STR mutation rates, which is especially seen for RM Y-STRs. Furthermore, we show how Y-STR mutation rates depend on Y-STR allele lengths and that this effect is linked with Y-SNP haplogroup background. We show that this finding has population specificity, in line with the population founder hypothesis on explaining Y-STR mutation rate difference between populations. The wider implications of this albeit sample size-limited study are two-fold. On one hand, by showing evidence for population differences in RM Y-STR mutation and male relative differentiation rates, our study highlights that the rates obtained from one population may not necessarily be transferable to samples from another population. For forensic practice this means that using RM Y-STR mutation rates established in for instance Europeans for interpreting RMplex casework results from a suspect of non-European paternal ancestry may be error-prone. On the other hand, our study calls for more RMplex population studies of suitably large (i.e., larger than used here) sample size to further study potential population effects on RM Y-STR mutation and relative differentiation rates, where Y-SNP haplogroups should be analyzed too to better understand the rate differences that may be observed in different populations. Ultimately, our study implies that there is a need to derive haplogroup-specific mutation and differentiation rates, which could lead to better interpretations when observing matching or closely-matching Y-STR haplotypes in forensic casework.

References

1. Kayser, M., *Forensic use of Y-chromosome DNA: a general overview.* Human Genetics, 2017. **136**(5): p. 621-635.
2. Ballantyne, K.N., et al., *Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications.* The American Journal of Human Genetics, 2010. **87**(3): p. 341-353.
3. Ballantyne, K.N., et al., *A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages.* Forensic Science International: Genetics, 2012. **6**(2): p. 208-218.
4. Ralf, A., et al., *Identification and characterization of novel rapidly mutating Y-chromosomal short tandem repeat markers.* Human Mutation, 2020. **41**(9): p. 1680-1696.
5. Ralf, A., et al., *RMplex: An efficient method for analyzing 30 Y-STRs with high mutation rates.* Forensic Science International: Genetics, 2021(55): p. 102595.
6. Neuhuber, F., et al., *Improving the differentiation of closely related males by RMplex analysis of 30 Y-STRs with high mutation rates.* Forensic Science International: Genetics, 2022: p. 102682.
7. Ballantyne, K.N., et al., *Toward Male Individualization with Rapidly Mutating Y-Chromosomal Short Tandem Repeats.* Human Mutation, 2014. **35**(8): p. 1021-1032.

8. Adnan, A., et al., *Improving empirical evidence on differentiating closely related men with RM Y-STRs: a comprehensive pedigree study from Pakistan.* Forensic Science International: Genetics, 2016. **25**: p. 45-51.
9. Boattini, A., et al., *Mutation rates and discriminating power for 13 rapidly-mutating Y-STRs between related and unrelated individuals.* PLOS One, 2016. **11**(11): p. e0165678.
10. Chen, Y., et al., *Mutation rates of 13 RM Y-STRs in a Han population from Shandong province, China.* Forensic Science International: Genetics Supplement Series, 2017. **6**: p. e346-e348.
11. Yuan, L., et al., *Mutation analysis of 13 RM Y-STR loci in Han population from Beijing of China.* International Journal of Legal Medicine, 2019. **133**(1): p. 59-63.
12. Zgonjanin, D., et al., *Mutation rate at 13 rapidly mutating Y-STR loci in the population of Serbia.* Forensic Science International: Genetics Supplement Series, 2017. **6**: p. e377-e379.
13. Zhang, W., et al., *Multiplex assay development and mutation rate analysis for 13 RM Y-STRs in Chinese Han population.* International Journal of Legal Medicine, 2017. **131**(2): p. 345-350.
14. Claerhout, S., et al., *Determining Y-STR mutation rates in deep-routing genealogies: identification of haplogroup differences.* Forensic Science International: Genetics, 2018.
15. Kelkar, Y.D., et al., *The genome-wide determinants of human and chimpanzee microsatellite evolution.* Genome Research, 2008. **18**(1): p. 30-38.
16. Brinkmann, B., et al., *Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat.* The American Journal of Human Genetics, 1998. **62**(6): p. 1408-1415.
17. Otagiri, T., et al., *Mutation analysis for 25 Y-STR markers in Japanese population.* Legal Medicine, 2021. **50**: p. 101860.
18. R Core Team, *R: A language and environment for statistical computing.* 2013.
19. Liu, J., et al., *The construction and application of a new 17-plex Y-STR system using universal fluorescent PCR.* International Journal of Legal Medicine, 2020. **134**(6): p. 2015-2027.
20. Ochiai, E., et al., *Y chromosome analysis for common surnames in the Japanese male population.* Journal of Human Genetics, 2021. **66**(7): p. 731-738.
21. Navarro-López, B., et al., *Phylogeographic review of Y chromosome haplogroups in Europe.* International Journal of Legal Medicine, 2021. **135**(5): p. 1675-1684.
22. Lang, M., et al., *Forensic characteristics and genetic analysis of both 27 Y-STRs and 143 Y-SNPs in Eastern Han Chinese population.* Forensic Science International: Genetics, 2019. **42**: p. e13-e20.
23. Liu, J., et al., *Development of a new 17 Y-STRs system using fluorescent-labelled universal primers and its application in Shanxi population in China.* Forensic Science International: Genetics Supplement Series, 2019. **7**(1): p. 95-97.

Supporting information

Supplementary Table S1: SNaPshot primer sequences and thermal cycling conditions

Y-SNP	Chromosome	GRCh38 position	Mutation	Forward PCR primer	Reverse PCR primer	PCR primer Concentration (µM)
CF-P143	chrY	12077161	G->A	CCCAACTTCGGTTTTTGGTA	GATTCTGGTCCCGTCTTTGA	0.2
C-M130	chrY	2866813	C->T	GATTTTGTTGGGTGGTGGTC	TGGCCAGCCTCTTATCTCTC	0.5
DE-M145	chrY	19555322	C->T	CCTCCCACTCCTTTTTGGAT	GCATACTTGCCTCCACGACT	0.2
D-M174	chrY	12842354	T->C	TCTCCGTCACAGCAAAAATG	GAAGGTCCTGGAGATGCAAA	0.2
N-M231	chrY	13357844	G->A	TCCTCTCCTCTCACAAAATGG	ATGGTGGCCAGAGTCTTTCA	0.2
O-M119	chrY	19600799	T->G	TGGGTTATTCCAATTCAGCA	CAAACCGCAGTGCTATGTGT	0.15
O-M122	chrY	19602788	A->G	CAGGCGATGCTGATATGCTA	TGCAAATGGTATGCAACTCAG	0.15
O-P186	chrY	7700527	C->A	GCTTCAGGAAGATCAACCCA	CATCACTGCATCTGGCTAACTT	0.5
O-P49	chrY	12384612	A->T	ACAGCACACTGGCAACTGAA	TCCATCCATGAACTGCTTCT	0.5
Q-M242	chrY	12906671	C->T	TTGTGCAAAAAGGTGACCAA	TTTCGCTTTAAGGGCTTTCA	0.2

Y-SNP	SBE primer ori	SBE primer	SBE primer length	Ancestral allele (color)	Derived allele (color)	SBE primer concentration (µM)
CF-P143	Forward	ttttttttttttGTGAAAATGTGTGGGGTT	30	G (blue)	A (green)	0.075
C-M130	Forward	tttttttttttttttttGCAATAAACCTTGGATTTC	36	C (black)	T (red)	1
DE-M145	Forward	tttttttttttttttttttttttttttGCTAAGGCTGGCTCT	42	C (black)	T (red)	0.075
D-M174	Forward	tttttttttttttttttttttttttttttttttCCTTCTGGAGTGCCC	48	T (red)	C (black)	0.2
N-M231	Forward	ttttttttttttttttttttttttttttttttttTTACTGTTTCTACTGCTTTC	54	G (blue)	A (green)	0.2
O-M119	Forward	ttttttttttttttttttttttttttttttttttttttttttCCAATTCAGCATACAGGC	60	T (red)	G (blue)	0.075
O-M122	Forward	tttttttttttttttttttttttttttttttttttttttttttttttttGATTTTCCCCTGAGAGC	66	A (green)	G (blue)	0.2
O-P186	Forward	tttttttttttttttttttttttttttttttttttttttttttttttttttttTGAAATTTTTCTGGGGAAC	72	C (black)	A (green)	0.4
O-P49	Forward	ttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttAGGAGGGGGAGAGAGA	78	A (green)	T (red)	1
Q-M242	Forward	ttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttAGGTGACCAAGGTGCT	84	C (black)	T (red)	0.2

Table S1 (continued)

Initial amplification			
Stage	**Time (min)**	**T ( )**	**Cycles**
Initial denaturation	10:00	95	-
Denaturation	00:15	95	35
Annealing	00:30	57	
Extension	01:00	72	
Final extension	07:00	72	-
Cooling	Forever	15	-
SBE reaction			
Stage	**Time (min)**	**T ( )**	**Cycles**
Initial denaturation	02:00	96	-
Denaturation	00:10	96	25
Annealing / extension	00:05	50	
Final extension	00:30	60	-
Cooling	Forever	15	-

Out of environmental considerations **Table S2** and **Table S3** belonging to this study were not printed with this chapter of the thesis. The digital files can be obtained from the publication at: https://doi.org/10.1016/j.fsigen.2022.102766

Chapter 6

Large-scale pedigree analysis highlights rapidly mutating Y-chromosomal short tandem repeats for differentiating patrilineal relatives and predicting their degrees of consanguinity

Arwin Ralf[1], Diego Montiel González[1] Dion Zandstra[1], Bram van Wersch[1], Nefeli Kousouri[1], Peter de Knijff[2], Atif Adnan[3], Sofie Claerhout[4,5], Mohsen Ghanbari[6], Maarten H.D. Larmuseau[7,8,9], Manfred Kayser[1]

[1] Department of Genetic Identification, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands
[2] Forensic Laboratory for DNA Research, Department of Human Genetics, Leiden University Medical Centre, Leiden, The Netherlands
[3] Department of Forensic Sciences, College of Criminal Justice, Naif Arab University of Security Sciences, Riyadh, Saudi Arabia
[4] Forensic Biomedical Sciences, Department of Imaging & Pathology, KU Leuven, Leuven, Belgium
[5] Interdisciplinary Research Facility Life Sciences, KULAK Campus Kortrijk, Kortrijk, Belgium
[6] Department of Epidemiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands
[7] Laboratory of Human Genetic Genealogy, Department of Human Genetics, KU Leuven, Leuven, Belgium
[8] ARCHES - Antwerp Cultural Heritage Sciences, Faculty of Design Sciences, University of Antwerp, Antwerp, Belgium
[9] Histories vzw, Gent, Belgium

Published in revised form in *Human Genetics*, October 2022
doi: https://doi.org/10.1007/s00439-022-02493-2

Abstract

Rapidly mutating short tandem repeats from the non-recombining part of the human Y-chromosome (RM Y-STRs) were suggested as genetic markers for differentiating patrilineally related men, which increases the power of Y-chromosomal approaches in forensic genetics, anthropological genetics, and genetic genealogy. Mutation and male differentiation rates of RM Y-STRs were previously estimated mostly from father-son pair data, while the required data from a broader range of patrilineal relatives is scarce. Here, we performed for the first time a large-scale pedigree analysis in 9,379 pairs of closely and distantly related men separated by one to 34 generations on 30 Y-STRs with increased mutation rates, including all currently known RM Y-STRs using the RMplex method. Part of these pairs were additionally genotyped by the state-of-the-art commercial Yfiler Plus kit including 25 Y-STRs. For 43 of the 49 Y-STRs analyzed in total, the newly obtained pedigree-based mutation rates were in line with the previous father-son based rates, while they differed significantly for six markers. Male relative differentiation rates based on the 30 RMplex Y-STRs were: 43%, 84%, 96%, 99%, and 100% for relatives separated by respectively one, four, six, nine, and twelve meioses. For these pairs, RMplex achieved an increase in male relative differentiation compared to Yfiler Plus of 217%, 165%, 109¬¬%, 21%, and 9%, respectively. Machine learning based models to predict the degree of patrilineal consanguinity based on Y-haplotypes yielded highly accurate and reasonably precise predictions when using RM Y-STRs. Matching haplotypes resulted in a 95% confidence interval of 1-6 meioses with RMplex compared to 1-25 with Yfiler Plus. Our results demonstrate the high value of RM Y-STRs for differentiating patrilineally related men of a broad range of relationship degrees, achieving in many cases individual identification, the lack of which has been the largest limitation of forensic Y-chromosome analysis until today.

Introduction

Short tandem repeats from the non-recombining part of the human Y-chromosome (Y-STRs) found their way to forensic research and casework application 30 years ago with the first described Y-STR [1, 2] and a few years later more Y-STRs followed [3]. The ability to obtain a male-specific STR profile from DNA mixtures that contain an excess of female DNA, such as commonly confronted with in cases of sexual assault involving a male perpetrator and a female victims, was instantly recognized [4] and let to the widespread use of Y-STRs in forensic casework within limited time [5]. Mutation rate studies of Y-STRs using father-son pairs [6] demonstrated that Y-STRs have similar mutation rates — generally in the order of one or a few mutations every 1000 generation per locus, as had been established earlier for their autosomal counterparts [7]. Such relatively low mutation rates explains why in the absence of recombination, male relatives typically share the same Y-STR haplotype. This haplotype sharing is advantageous when conducting genetic genealogical research [8]. For example, a non-matching haplotype may indicate a discrepancy between the biological pedigree structures and legal family records [9], while shared haplotypes can confirm the biological validity of such records. However, the general lack of Y-haplotype variation within patrilineal relatives also poses limitations to genetic genealogy; for example, low precision when estimating the level of relatedness based on two similar haplotypes [10]. Due to their conservation over time, Y-haplotypes can also be used in anthropological genetics, e.g., to gain understanding in population substructure [11], to trace migration patterns [12], or to detect founder effects [13].

In forensic genetics, a match between a standard Y-STR haplotype of a male suspect and that of a crime scene sample means that the crime scene sample could have originated from the male suspect. However, a matching standard Y-STR haplotype could also have originated, with the same statistical evidence, from any of his close or distant paternal relatives [14], reflecting a limitation. Hence, it is up to tactical police investigation to establish, by excluding all of his paternal male relatives, that the matching suspect was indeed the likely sample donor. Unequivocally excluding all male relatives of the matching suspect that share the same Y-STR haplotype becomes increasingly difficult the more of such relatives exist. The relatively low number of Y-STRs and the high haplotype resemblance within various Y-SNP based haplogroups due to radiation, additionally led to a relatively high number of shared Y-STR haplotype between unrelated males (identity by state, IBS) especially with the earlier Y-STR kits [15, 16]. Recently, by continuously increasing the number of Y-STRs in the next generation of commercial Y-STR kits, the IBS problem became smaller and paternal lineage identification gained specificity. However, because most Y-STRs included in commercial kits have relatively low mutation rates of a

few mutations in 1000 generations per locus, Y-STR haplotype sharing between related men remains a major problem of these kits.

A turning point was marked by the findings of a large-scale Y-STR mutation rate study regarding both the number of Y-STRs (a total of 186) and the number of father-son pairs confirmed with autosomal DNA (close to 2000) [17]. In that study [17], 13 Y-STRs with remarkably high mutation rates, exceeding 10^{-2} mutations per generation (mpg), were identified and termed rapidly mutating Y-STRs (RM Y-STRs) [18]. These and subsequent studies demonstrated that RM Y-STRs strongly increase the differentiation of paternally related males compared to standard Y-STRs because of their increased mutation rates [19]. Moreover, RM Y-STRs were also shown to improve the differentiation of unrelated males compared to AmpFLSTR™ Yfiler™ PCR Amplification Kit, the state-of-the-art commercial Y-STR testing kit at that time [20]. As a result of these scientific developments, industry picked-up on these findings and included some (but not all at the time known) RM Y-STRs in their next generation commercial Y-STR kits such as the Yfiler™ Plus PCR Amplification Kit (in the following referred to as Yfiler Plus) [21] and the PowerPlex Y23 System [22].

Recently, more RM Y-STRs were discovered that further improved the male relative differentiation rates and further increased the advantage over standard Y-STRs in differentiating paternally related men [23]. Subsequently, a new genotyping method named RMplex was developed to analyze a total of 30 Y-STRs with increased mutation rates including all 26 currently known RM Y-STRs [24]. Most recently, a father-son pair study involving ~500 pairs [25] demonstrated that RMplex is highly effective and allows to differentiate fathers from their sons in over 40% of the cases and, albeit based on a more limited dataset, 62% of brother pairs. In comparison, the current state-of-the-art commercial Y-STR kit Yfiler™ Plus achieved differentiation in only 13% of the father-son pairs and 33% of the brother pairs in the same samples [25]. However, data on how these 30 RMplex Y-STRs differentiate more distantly related males is lacking completely thus far as empirical studies in more distantly related males are not available as of yet.

Up to now, knowledge on mutation rates and male relative differentiation rates of RM Y-STRs was mostly established in father-son pair studies [17, 20, 23, 26-28], which in principle only allow for the estimation of how closely related males can be differentiated. Pedigree studies, on the other hand, have the advantage that a broad range of male relationships can be studied and a large number of meiotic divisions can be covered by analyzing only a small number of male samples. This makes such pedigree studies more efficient in reaching the large numbers of meioses needed to establish reliable mutation rate estimates [29-31]. Mutation rates estimated from pedigree studies

come, however, with more uncertainties than those from father-son pair studies, which needs to be considered. On the other hand, for investigating male relative differentiation, pedigree studies have a clear advantage over father-son pair studies because they include both closely and distantly related males. The more men that can be genotyped and the deeper the pedigrees are rooted, the more types of distantly related males are available.

Here, for the first time, we performed a large-scale pedigree study on RM Y-STRs by analyzing 1,793 males belonging to a total of 403 pedigrees from three cohort studies of diverse bio-geographic ancestries, allowing for a total of 9,379 pairwise comparisons of closely and distantly related men separated by one up to 34 generations. We genotyped 30 Y-STRs with increased mutation rates, including all currently known RM Y-STRs, using the RMplex genotyping method. For most of the relative pairs, we additionally genotyped the current state-of-the-art commercial Yfiler Plus Kit, allowing the direct comparison of the results obtained with both Y-STR kits. We estimated male relative differentiation rates for all degrees of relationships based on RMplex and Yfiler Plus. Moreover, we estimated the mutation rates of the total of 49 Y-STRs and compared them with previous mutation rate estimates established from father-son pairs. Finally, we developed machine-learning based models based on simulated data to predict the degree of patrilineal consanguinity based on differences in the Y-STR haplotypes of two related males, and validated them using the empirical data obtained in this study.

Results

Mutation rates

In this study, three cohorts were analyzed, these cohorts consist of pedigrees characterized by different depths of rooting, different sample sizes, different demographic characteristics, and different biogeographic ancestries. The pedigree-based mutation rates were estimated per each cohort separately and for all three cohorts combined (Table S1). For the Yfiler Plus specific loci, only Cohort 1 was included, as the individuals from the other two cohorts had not been genotyped for that assay. The pedigree-based mutation rates were compared to father-son based consensus mutation rate estimates, which were recently published based on multiple father-son based studies [25] (Table S1). For the vast majority of Y-STRs (i.e., 43 of the 49 Y-STRs analyzed in total with both kits), the obtained pedigree-based mutation rates were coherent with the father-son based mutation rates previously established for these markers . Six Y-STRs showed significant differences between the two ways of estimating mutation rates: DYF1000, DYF403S1a, DYS612,

DYS1013, DYS442 and DYS448 (Table S1). For three of those i.e., DYS1000, DYF403S1 and DYS612, the pedigree-based mutation rate estimates were significantly higher than the father-son based rates (p-value 0.001-0.018). Notably, all these three Y-STRs showed high mutation rates in absolute sense (i.e., >0.02 mpg), whereas the three Y-STRs with significantly lower (p-value 0.009-0.047) pedigree-based estimates i.e., DYS1013, DYS442 and DYS448 showed lower absolute mutation rates (i.e., <0.005 mpg). Differences in mutation rate estimates could also be found between the three different cohorts (Figure 1, Table S1), although the overall trends appeared rather consistent across the total pedigree dataset. Notable cohort specific outliers were DYF1000, DYF387S1 and DYS518, which showed remarkably high mutation rates in Cohort 3, which consisted of Pakistani males. On the other hand, Cohort 2, which is characterized by its deep rooting, showed a markedly lower mutation rate estimate for DYS724 compared to the other pedigree cohorts and the father-son based reference. Figure 1 presents the data for all cohorts for the 30 RMplex Y-STRs, while the data for all 49 Y-STRs, including the Yfiler Plus loci, are given in Table S1.

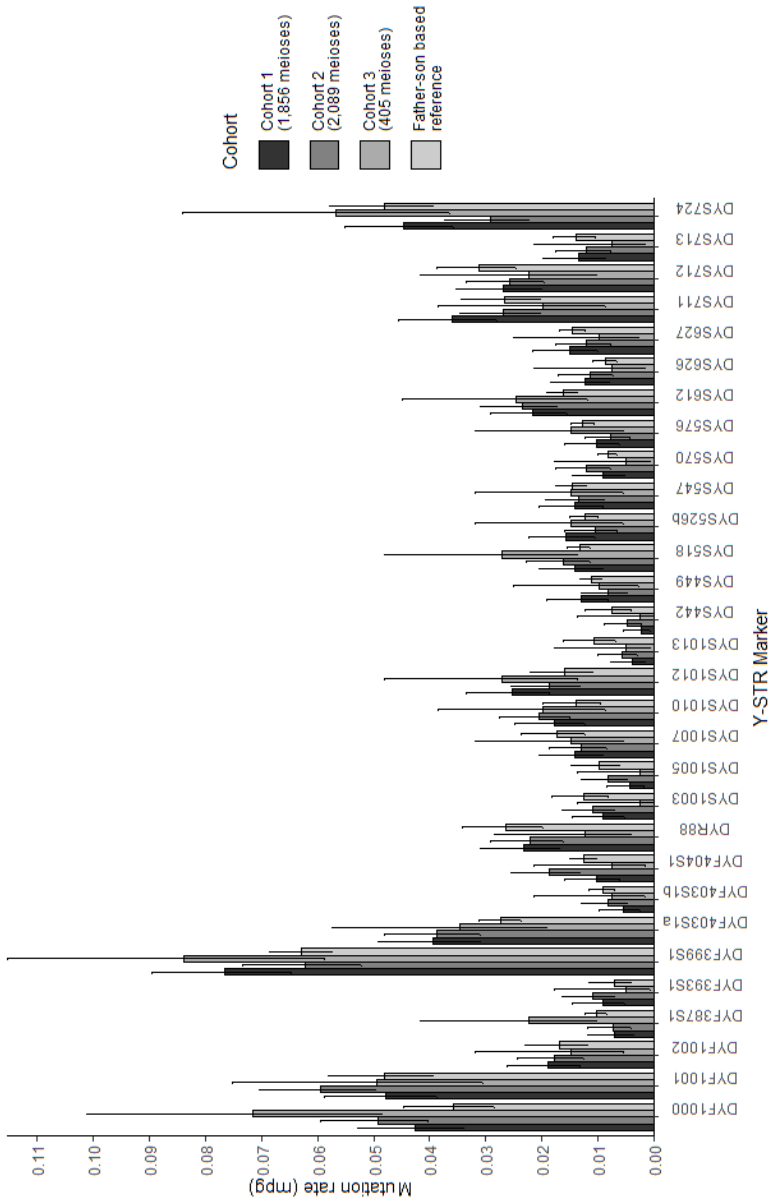


Figure 1: Pedigree-based mutation rate estimates for 30 RMplex Y-STRs from three cohorts as well as the father-son based reference consensus estimates (based on 2,025-12,387 meioses per Y-STR) from a recent study [25]. The error bars represent the 95% Clopper-Pearson intervals.

Male relative differentiation rates

The male relative differentiation rate of a given set of Y-STRs refers to the rate at which a given pair of paternally related males (e.g., brothers, or first cousins) can be discriminated from each other by an allelic difference in at least one Y-STR marker. In contrast to mutation rates, male relative differentiation rates are of direct practical forensic relevance as they provide investigators with an expectation on the chance of being able to distinguish paternally related males depending on their degree of relationship. Efficient male relative differentiation is also of high importance for genetic genealogical research, as it can help to place males correctly into larger pedigrees.

By taking advantage of the deep-rooted nature of a part of the pedigrees analyzed here, we were able to establish differentiation rates for male relatives separated from one meiosis (i.e., separated by one generation: a father-son pair) up to 34 meioses. All RMplex data are presented in Table S2. Overall, by combining the results from all three cohorts, RMplex achieved a differentiation rate of 43,3% for males separated by one meiosis, while males separated by two meioses (i.e., brothers and grandfather / grandsons) were differentiated in 66% of the cases. Moreover, relative differentiation for males separated by six or more meioses was over 95%, and male relatives that were twelve or more meioses apart were differentiated 100% of the time. Notably, the sample size of male relatives separated by one to thirteen meiosis was rather large with 334 to 966 pairs, while for those fourteen or more meiosis apart was markedly smaller (i.e., less than 100 pairwise comparisons). On one hand, reliability of the estimates increases with sample size. On the other hand, since our applied criterion of male relative differentiation was an allelic difference at one Y-STR at least, and the change of mutations occurring increases with the number of separating meioses (Table S2), it can be expected that also with a larger sample size, males separated by more than 13 meioses will be differentiated in 100%, or nearly 100%.

For Cohort 1, we describe the results in more detail because this cohort contains pedigrees that include a large number of different degrees of relatives, especially for previously understudied distantly related males up to 13 generations apart and because Cohort 1, additionally to RMplex, also has Yfiler Plus data available which allows for direct comparison (Figure 2, Table S3). This comparison highlighted that RMplex showed to be

far more superior to Yfiler Plus, in regards of the differentiation of both closely and more distantly related males (Figure 2). Using Yfiler Plus, only 10% of the father-son pairs were differentiated, compared to 44% with RMplex. Combining the two assays only led to a marginal increase to 45% compared to RMplex alone. The differentiation rates increased with the number of meioses between two related males (Figure 2), as was expected given the independent probabilities with which mutations occur during every meiosis that separates two relatives. Because at least one mutation, i.e., one allelic difference, is sufficient to discriminate between two males, the noted increase appeared more dramatic in the relative pairs separated by fewer number of meioses than in those separated by larger number of meioses (Figure 2). RMplex was able to differentiate over 95% of the male relatives separated by six meioses, while only 42% of such relatives were separated with Yfiler Plus. Complete differentiation of all relative pairs was achieved in men separated by twelve and more meioses using RMplex, by ten and more meioses using the combined assays, and never up to the thirteen meioses with Yfiler Plus. Yfiler Plus had a maximum differentiation rate at 90% in males separated by 13 meioses, which was below the differentiation rates already achieved with RMplex in males separated by five meioses.

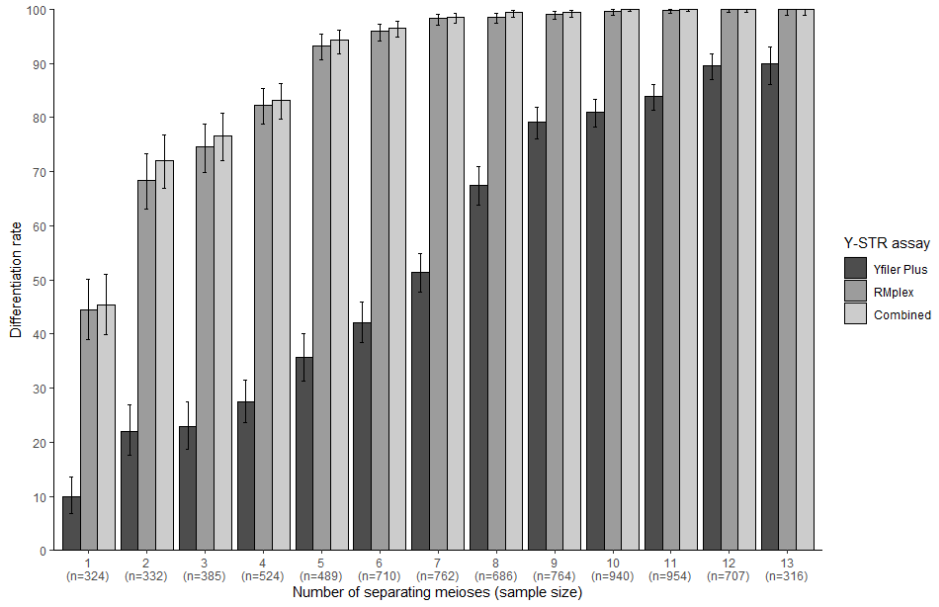


Figure 2: Male relative differentiation rates obtained from Cohort 1 pedigrees for RMplex (30 Y-STRs), Yfiler Plus (25 Y-STRs), and both assays combined (49 Y-STRs) for pairs of males related by 1 to 13 meioses. The error bars represent the 95% Clopper-Pearson intervals. Male relative differentiation is defined as a pair having at least one (but not excluding multiple) allelic differences.

To exemplify how the differences in differentiation rate between the marker sets and assays affect the ability to differentiate individuals within a given pedigree, Figure 3 shows two examples of pedigrees from Cohort 1. Figure 3 (a-c) each shows a total of 21 genotyped individuals; using Yfiler Plus (Figure 3a), a total number of five unique haplotypes was observed, including a single haplotype that uniquely identified a single individual, whereas the most commonly observed haplotypes was shared by eleven of the 21 genotyped males. In the same pedigree using RMplex (Figure 3b), a total number of fifteen haplotypes was observed, of which six were uniquely attributed to single individuals, while the most commonly observed haplotype was shared by only three of the genotyped individuals. By combining both assays (Figure 3c), a total of 17 haplotypes were observed of which seven could be attributed to single individuals, and the most common haplotype was shared by only two individuals. Figure 3 (d-f) shows a similar pattern in a different pedigree, where using Yfiler Plus (Figure 3d) only two haplotypes were observed, while with RMplex (Figure 3e) eleven different subgroups of relatives were seen, without any further improvement when both assays were combined (Figure 3f). Both pedigrees exemplify the strongly improved male relative differentiation achieved by RMplex compared to Yfiler Plus for male of different degrees of paternal relationships.

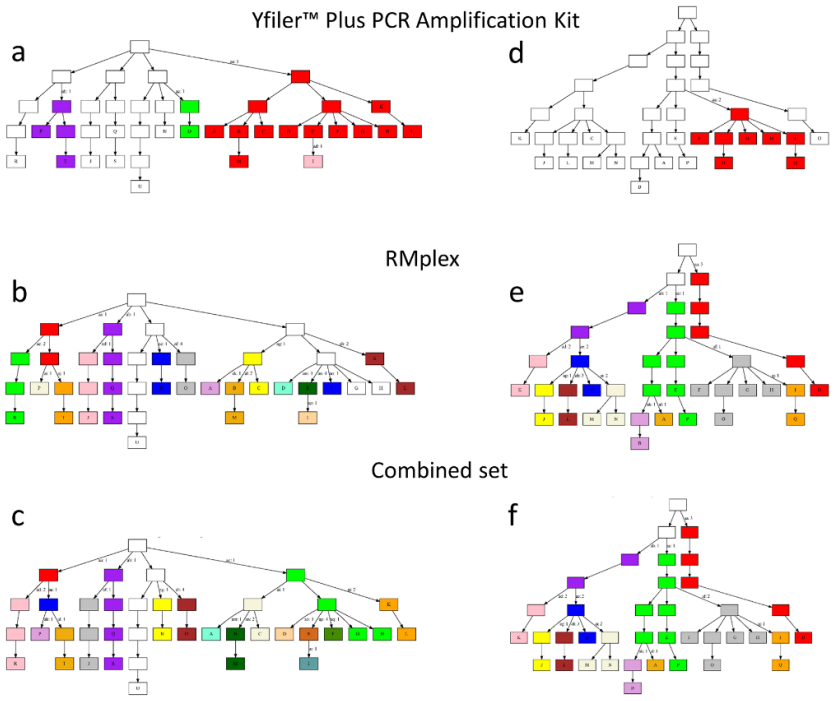


Figure 3: Male relative differentiation in two example pedigrees using Yfiler Plus (a, d), RMplex (b, e), and both assays combined (c, f). The different colors indicate unique haplotypes different from the inferred ancestral haplotype shown in white. The nodes with labels indicate individuals that were genotyped; individuals with unlabeled nodes were unavailable for genotyping. The colors in the unlabeled nodes indicate hypothetical haplotypes as the mutations could have occurred in any patrilineal ancestor that shares the color of the genotyped individual(s). The letters on the labels next to the arrows correspond to specific (sets of) mutations observed, whereas the numbers reflect the total number of mutational steps.

Prediction of patrilineal consanguinity

Next, we investigated if the observed differences in Y-STR genotype data between two related males can function as a reliable predictor for the degree of patrilineal consanguinity between those two males. To this end, we employed a machine learning approach to develop models, which were trained on simulated data, that can predict the degree of patrilineal consanguinity based on the observed Y-STR allelic differences, i.e. mutations, between two related males, for RMplex and Yfiler Plus data separately, as well as for the combined dataset. Figure S1 shows the results of those models for the scenario where no allelic differences were observed, i.e., a matching haplotype between the relatives, which would indicate a close relationship, particularly when many RM Y-STRs are included as with RMplex. Indeed, the 95% confidence interval for RMplex ranged from one to six meioses. For Yfiler Plus, however, the 95% confidence interval was much wider, with one to 25 meioses, demonstrating a larger uncertainty about the relationship in the case of a matching Y-STR haplotype. When combining both assays, the 95% interval remained one to six meioses; however, the combined probability (i.e., the sum of the probabilities obtained for each distance included in the interval) slightly increased from 95.5% with RMplex to 96.3% with both assays combined. Y-STR mutations are highly stochastic, as indicated by the high variance shown in Figure S2. On average, the number of observed allelic differences increases the more distant the paternal familiar relationship is. As expected, for RMplex this trend was seen a lot stronger than for Yfiler Plus; while at the same time the variance observed in RMplex was also larger. Generally, there was a strong overlap in the distribution of number of observed mutations between different meiotic distances, especially those in close proximity of one another.

To empirically demonstrate that indeed Y-STRs with a high mutation rate such as RM Y-STRs are more suitable for the purpose of predicting patrilineal consanguinity compared to standard Y-STRs with lower mutation rates, the newly developed models for Yfiler Plus, RMplex, and the two assays combined were empirically tested on all pairs of paternally related men of different degrees. At this end, we used the data from Cohort 1, because of the reasonably large sample size per each degree of relatedness being

available in this cohort for male relatives separated by one to thirteen meioses ranging from 316 to 954 pairwise comparisons. Therefore, these generational groups were evaluated separately. Additionally, all pairs included in the cohort, including those separated by more than 13 meioses, were analyzed as a whole. The two most critical characteristics for predicting the degree of paternal relationship from the Y-STR data were evaluated: prediction accuracy (i.e., the percentage of pairs of which the true value fell within the prediction intervals) and precision (i.e., the size of the prediction intervals). As could be expected, the precision of Yfiler Plus fell short of that of RMplex as indicated by the relatively large prediction intervals (Figure 4), and the precision of the two assays combined was slightly higher than that of RMplex alone. Obviously, the prediction intervals (Figure 4) and the prediction accuracy (Figure 5) increase when higher confidence levels are considered; increased prediction intervals equal reduced precision. Another trend that became evident is that the size of the prediction intervals also increased in more distant relationships (Figure 4). With regards to accuracy Yfiler Plus resulted in slightly more accurate predictions compared to RMplex and the combined assays. When looking at the overall prediction, i.e., including all levels of relationship, Yfiler Plus resulted in correct prediction in 93.0%, 95.8%, and 98.4%, for predefined confidence levels of 85%, 95%, and 99%, respectively. RMplex resulted in accurate prediction in 86.7%, 95.5%, and 98.6% for the same confidence levels, respectively; while the two assays combined predicted accurately in 87.1%, 95.3%, and 98.5%, respectively (Figure 5). The prediction accuracy was not constant among the different number of separating meioses, the accuracy of our models appears to be somewhat reduced in the proximity of nine meioses (Figure 5). The models described here and a number of additional models for different (combinations of) Y-STRs kits that have not yet been empirically validated can be used through a web user interface on: https://ystr.erasmusmc.nl.

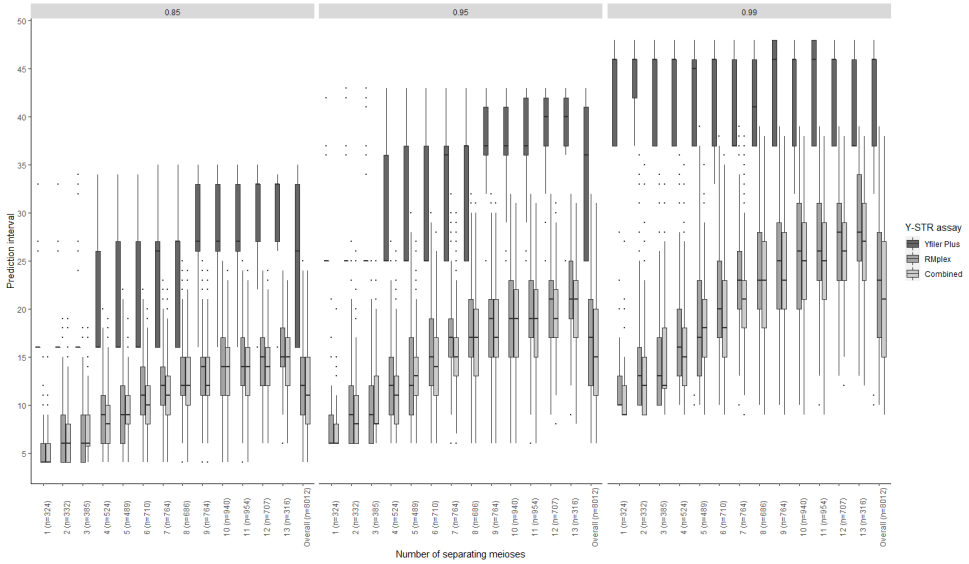


Figure 4: Boxplots showing the distribution of the prediction intervals of the three different machine learning models trained to predict the degree of patrilineal consanguinity based on the observed mutations between pairs of paternally related males using Yfiler Plus, RMplex and both assays combined using three different predefined levels of confidence (85%, 95% and 99%).

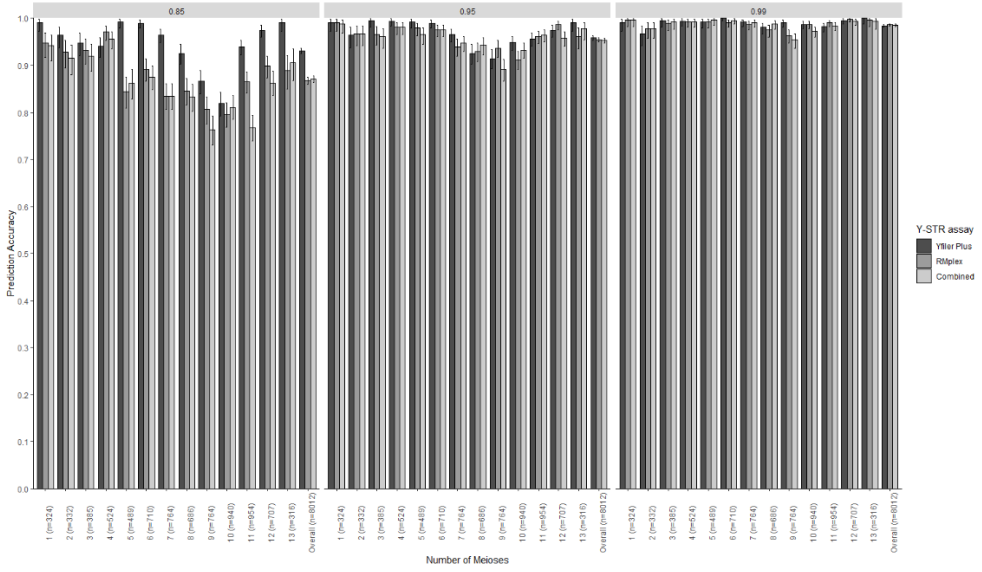


Figure 5: The accuracy of three different machine learning models trained to predict the degree of patrilineal consanguinity based on the observed mutations between pairs of paternally related males using Yfiler Plus, RMplex, and both assays combined using three different predefined levels of confidence (85%, 95% and 99%). The error bars represent the 95% Clopper-Pearson intervals.

To put the performance of our newly developed machine learning-based models (MLM) in perspective, we compared the results to two intensively studied models to describe STR variations: the infinite alleles model (IAM) and the stepwise mutation model (SMM). All three models were evaluated by testing the same set of 1000 randomly selected pairs of paternally related men from all three cohorts. Notably, IAM outperformed the other two models both in regard of prediction accuracy (Figure 6a) and precision (Figure 6b); SMM, in turn, was the least well performing model out of the three. The accuracy of IAM was significantly higher than that of SMM (Fisher's exact p-value: 0.0204); the difference between IAM and MLM was not significant (p-value: 0.1143), nor was the difference between SMM and MLM (p-value: 0.5376). All three models delivered an accuracy >95% (Figure 6a), which was expected as the 95% confidence intervals were used by all models. To learn more about the nature of the prediction errors resulting from each of the three models, Venn diagrams were used for the total number of errors (Figure 6c), the overestimations (Figure 6d), and the underestimations (Figure 6e). Overestimations were the most common type of prediction errors in each of the three models. Some pairs consistently lead to errors regardless of the model that was used; SMM and, to a slightly lesser degree, MLM overestimated the number of generations more often than IAM (Figure 6d). Notably, SMM showed the lowest number of underestimations and in cases where it did, it was consistent with the other two models (Figure 6e).

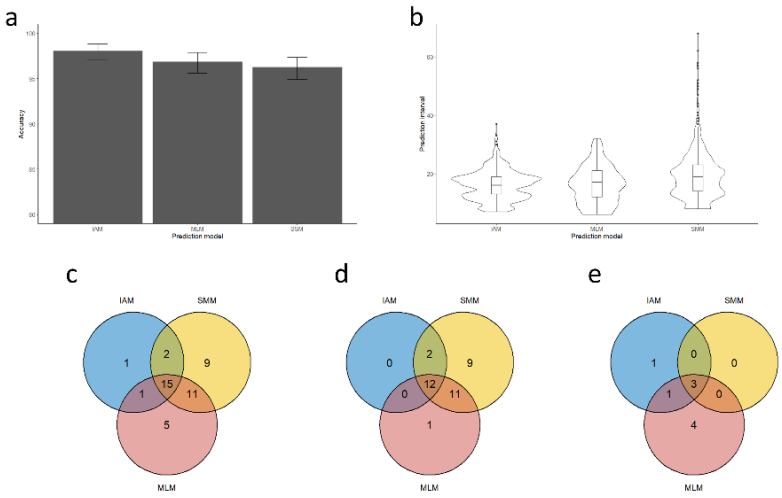


Figure 6: The performance of three different models to predict the degree of paternal consanguinity based on RMplex Y-STR data: Infinite Allele Model (IAM), Single Mutation Model (SMM) and the newly developed Machine Learning-based model (MLM) using 95% confidence intervals. The performance was assessed by the accuracy (a), the precision (b). The recurrence of errors was further evaluated by using Venn diagrams showing the total number of errors (c), overestimations (d), and underestimations (e). Numbers in c-e reflect the total numbers out of the total of 1000 pairs that lead to incorrect predictions.

Discussion

Mutation rates

Previous RM Y-STR mutation rate studies mostly focused on father-son pairs. The advantage of such studies is that the degree of relationship, i.e., the number of separating meioses is known with certainty, e.g., because only father-son pairs with paternity confirmed with autosomal DNA from analyzing complete trio cases were used. Hence, from an observed Y-STR allelic differences between a father and his biological son, it can safely be concluded that a mutation had occurred. The disadvantage is that, unless very large numbers of father-son pairs are analyzed, the statistical power is low. Limited statistical power leads to limited reliability of the obtained mutation rate estimates. In addition to the costs and labor associated with typing such large number of samples,

sample availability is also a limiting factor that needs to be overcome to perform accurate father-son pair based mutation rate studies.

Estimating mutation rates from pedigrees, on the other hand, comes with the advantage that, depending on the deep-rooting structure of the pedigree, many meiosis can be covered by analyzing only a restricted number of males. Thus, pedigree studies typically reach larger numbers of meioses, which theoretically allows for more reliable mutation rate estimates. The cost-effectiveness of using especially deep-rooting pedigrees can be striking. For example, in Cohort 2 of this study, a total of 2,089 meioses were covered by analyzing only 265 individuals. To cover the same number of meioses using the father-son based approach would require genotyping almost 4,200 individuals; ergo, 15-fold increased genotyping efforts and resources. However, the reliability of pedigree-based studies can be hampered by uncertainties. One of such uncertainties would be the presence of extra-pair paternity events, which are estimated to occur at a frequency of ~1% in human populations [9, 32]. Typically, individuals who are not biologically related to the others in the pedigree can be easily detected as the observed genotypic variation between the pair is larger than can be explained by mutations alone. However, if the biological father was a paternal relative of the legal father — hence, the individual belongs to the pedigree but its place in the pedigree is different than assumed by the pedigree records — this can be missed. Moreover, estate/parish records and genealogical analyses may contain other flaws. This could lead to misinterpretations of the number of generations that separate two males in a pedigree. Another uncertainty when estimating Y-STR mutation rates from pedigree data is the possibility of parallel mutations, and back- and forward mutations [23, 33, 34], which both cannot occur between a father and his son because they are only separated by one meiosis. In the case of Y-STRs with complex repeat structures, sequencing instead of fragment length analysis may provide additional information that can differentiate the scenarios from one another [33]. Here, we decided to always assume that when no difference in genotype was observed no mutation had occurred, and when multiple mutational steps were observed, (with the exception of father-son pairs) we assumed that this was the result of multiple single step mutations. As the effects of the errors that may result from these two assumptions will be in opposite directions, they will become at least partially mitigated.

In our study, 43 of the 49 Y-STRs we analyzed in total had pedigree-based mutation rate estimates that were not significantly different from the previous father-son based reference mutation rates that were established from fairly large numbers of father-son pairs. Only six out of the 49 Y-STRs analyzed showed significantly different mutation rate estimates from the current pedigree data, compared to the previously obtained father-son based reference mutation rates. It is difficult to know the exact reason for

these differences, which can be intrinsic to the methodology employed, or not. It is remarkable, that the three Y-STRs with higher mutation rate estimates in the current study all showed markedly higher absolute mutation rates compared to the three Y-STRs that displayed lower mutation rates in this study. The different mutation rate estimates could also be caused by stochastic effects, or could be a result of the different populations that were being used [31]. Both the reference mutation rates based on the previous father-son pair studies [25] and the mutation rates estimated here from the pedigrees were based on a large numbers of meioses (i.e., in the thousands), which together with the increased mutability of 30 of the 49 Y-STRs analyzed here, makes the overall statistical power of the analyses large for both approaches. Hence, stochastic effects driven by limited sample size may not play a major role to explain the differences at these six Y-STRs. Future studies may shed more light on which estimates form the best approximation of the locus-specific mutation rates and what are the merits and demerits of each of both approaches.

Male relative differentiation

Here we performed the most comprehensive study into male differentiation rates based on Y-STRs available to date, regarding the number of Y-STRs, the number of male relatives, and the number of degrees of paternal relationships we considered. Previous studies mostly focused on less Y-STRs and only used close relatives such as father-sons and brothers, or only on a limited number of generations [18-20, 25, 27, 29, 35-37]. The father-son differentiation rates of 10%, 44%, and 45% for Yfiler Plus, RMplex, and both assays combined, respectively (Figure 2), that we obtained in the current pedigree study is comparable to the father-son differentiation rates of 14%, 42%, and 48%, respectively, previously established from father-son pairs for the same marker sets [25]. By deriving male relative differentiation rates solely from mutation rate estimates, thereby describing the theoretical expectations of male relative differentiation, another study estimated based on a slightly reduced set of 26 of the 30 RMplex Y-STRs used here, rates of 44%, 69%, 83%, and 90% for male relatives separated by one to four meioses, respectively [23]. In the current pedigree study, we obtained differentiation rates from empirical data of 43%, 66%, 76%, and 84%, respectively (Table S2), using RMplex. Hence, the previous theoretical rates closely agree with the empirically derived rates for relatives separated by one and two meioses, while for those separated by three and four meioses, the theoretical rates represent slight overestimations. The same previous study [23] also hypothesized based on differentiation rates estimated from mutation rates that male

relative differentiation rates of 99% would be achievable from eight meioses onwards. In the current pedigree study, we empirically showed that the 99% differentiation rate was reached with RMplex from nine meioses onwards, closely agreeing with the previous theoretical expectation. Moreover, our study is the first that demonstrates male relative differentiation in appreciable numbers for distant relatives separated by more than two meioses for the full set of 30 RMplex Y-STRs, previously only father-son pairs and a limited number of brothers were described [25]. Male relative differentiation of males separated by three to four meioses were only available for a subset of 13 RM Y-STRs [19], and reliable data (i.e., with sufficient sample size) about Y-STR differentiation of males separated by more than four meioses was lacking completely. Overall, RMplex did fulfil its promise of delivering male relative differentiation with an unprecedented efficiency for all degrees of paternal relationships, as demonstrated.

The differentiation rates can provide forensic investigators with an expectation about the evidential value of a Y-STR haplotype match. Historically, the strongest value of Y-STRs in court cases has been to exclude a male suspect as being the donor of a crime scene stain. While, conversely a fully matching Y-STR haplotype was considered a non-exclusion. The state-of-the-art method to determine the value of a non-exclusion is through the use of population frequency databases such as YHRD [38]; the more frequently a Y-STR haplotype is observed in such databases, the lower the evidential value is regarded [39]. Additionally, more complex statistical methods can be employed to estimate the frequency of the haplotype in the population even when the observed haplotype is not found in the database [40, 41]. Although the database-derived population frequency approach seemed to work well with the older generation of Y-STR kits, the more recent versions, like Yfiler Plus that contain more Y-STRs, including a limited number of RM Y-STRs, have a much larger discrimination capacity and haplotype diversity, which results in the need for much larger databases. However, even in large frequency databases it can be expected that there will be many singletons (i.e., haplotypes observed only once in a population), or haplotypes that are not present in the database at all, because of its limited size relative to the whole population and given the diversity of the haplotypes. Hence, determining population frequencies of haplotypes becomes increasingly challenging the more Y-STRs are analyzed. This effect would become even more pronounced for Y-STR kits containing large numbers of RM Y-STR. The differentiation rates obtained in this study show that, generally, only paternally related males separated by just a relatively low number of meioses share Y-STR haplotypes when using RMplex.

Therefore, instead of relying on statistical methods, to calculate the evidential value of a match, RMplex – due to the high differentiation rates which are found even among close relatives – has the potential to exclude (close) male relatives of the real perpetrator by revealing non-matches. In forensic cases where commercial Y-STR kits such as Yfiler Plus or PowerPlex Y23 revealed a match between the male suspect and the DNA from a crime scene stain, RMplex can be applied to further investigate this match. If for instance the suspect was, in fact, not the crime scene sample donor, but instead one of his male relatives was, RMplex has a very good chance of demonstrating this by showing a non-match with the suspect. This chance is very high for more distant male relatives of the suspect, but also fairly high for even his closest relatives as our data demonstrates. In high-profile cases it may even be worthwhile to characterize all known living paternal relatives of the male suspect to further limit the number of potential contributors to the crime scene stain. Although this alone would not solve the case, it would free wrongly suspected men from further investigation. By this approach the potential contributors to a crime-scene sample could ideally be reduced to a single man, and in most cases to only about a handful of close male relatives, as exemplified in Figure 3. With such a small pool of potential contributors, additional investigative techniques could in many cases identify which of those males was the true contributor to the crime scene trace.

Moreover, the high differentiation rates of RM Y-STRs and RMplex provide a solution to genetic genealogist when it comes to males. With the tools that typically are at their disposal, i.e., Y-STRs, Y-SNPs, and autosomal DNA markers, it can be challenging to place an individual in the right position within a pedigree. RM Y-STRs, however, as can be asserted from the examples in Figure 3, would allow to localize an individual's position in a given pedigree with more precision. Furthermore, in anthropological genetics, in particular in population influenced by strong founder effects, male differentiation using RM Y-STRs can uncover population substructure when standard Y-STRs cannot because of high levels of homogeneity in the population. Lastly, the increased ability of RM Y-STRs to differentiate relatives may also be suitable to study recent migration events.

Prediction of the degree of patrilineal consanguinity

Our results show that despite the stochastic nature of Y-STR mutations, it is feasible to predict the degree of patrilineal consanguinity of two males within a reasonably narrow

range solely based on the number of observed Y-STR variations . We also showed that a higher precision (i.e., more narrow confidence intervals) could be achieved by analyzing Y-STRs with higher mutation rates, demonstrating the superiority of RMplex also for this purpose, over commercial Y-STR kits such as Yfiler Plus. This latter finding is in agreement with a previous study that also found RM Y-STRs to deliver more precise estimations of the time since the most recent common ancestor (TMRCA) for other than forensic purposes [30]. Furthermore, we have shown that it is feasible to develop prediction models based on simulated Y-STR mutation data. The accuracy of the predictions based on our empirical data was largely in agreement with the expected accuracy based on the simulated data. The implication of these results is that such models can easily be developed for other sets of Y-STRs, given that the mutation rates of all markers in such a kit are known. In addition, multiple models could be built for the same sets of Y-STRs, based on different mutation rate estimates, for example if it is shown that the marker-specific mutation rates strongly differ in the population of interest. This method of investigation may become more precise over time as the number of addressable and well-characterized Y-STRs increases, for example by using massively parallel sequencing-based methods for data generation [42].

In forensic genetics as well as in genetic genealogy, it is possible to encounter fully, or nearly matching Y-STR haplotypes, while other knowledge about the relationship of the two matching males is unavailable. In the past, when only a small number of Y-STRs were typed, there was a reasonable chance that two individuals shared a haplotype while not being related to any meaningful degree due to IBS [15]. The more recent versions of commercial Y-STRs kits, contain more markers, thereby strongly reducing the probability of a haplotype match due to IBS. However, even when using currently available, improved, commercial Y-STR kits, such as Yfiler Plus, matching haplotypes can be detected in related men that are descendants of a male that lived many generations ago, as we demonstrated here (see Figure 3). This also became apparent in Figure S1a, where it was shown that the 95% confidence interval for a fully matching Yfiler Plus profile ranges from 1 to 25 meioses. Hence, even if a full Yfiler Plus haplotype match was found between two males, this may only indicate that they share a common ancestor that dates back more than ten generations, i.e., several hundreds of years. In comparison, for RMplex the 95% interval ranges from one to six meioses. In cases when two males show a matching Yfiler Plus profile as the result of a distant common ancestor, RMplex would likely show multiple allelic variations and reflect the more distant relationship in the resulting prediction. We provide empirical evidence for this with our study in general by seeing improved relative differentiation rates with RMplex compared to Yfiler plus, but also in example pedigrees

showing many male relatives with matching Yfiler Plus haplotypes that can be separated by RMplex. This information can, for instance, be valuable to genealogists trying to understand the relevance of an unexpected Y-STR match.

In our study, we found that the infinite alleles model (IAM) outperformed both stepwise mutation model (SMM) and the novel machine learning model (MLM) that we have introduced in the present study, although the differences were not striking. These results contradict a recent study by Claerhout *et al.* [43] which found SMM to outperform IAM, while in that study both methods delivered an accuracy that was well below the accuracy we found in the present study. This previous study also proposed a new method that was found to deliver more accurate results than IAM and SSM [43]. Unfortunately, we were unable to apply this newly proposed method, possibly due to the large number of RM Y-STRs included in our study leading to technical errors, potentially related to memory issues. Therefore this method was not included in the comparison made here. Another study [30], however, found IAM to be more accurate than SSM, which is in accordance with our results. The described accuracy in this latter study was higher than that described in the study from Claerhout *et al.* [43] for both models, but still lower than the accuracies that were achieved for IAM and SSM in the current study. A potential explanation for the reduced accuracy that was observed in both previous studies may be that both studies included more distantly related males, i.e., deep routed pedigrees; whereas the randomly drawn pairs in the present study predominantly were separated by one to thirteen meioses, as over 95% of our pairs were separated by meiotic distances in that range. Our data suggests that all models are valid and provide accurate predictions according to their confidence intervals. However, in our study IAM demonstrated a slightly better accuracy. It may be that the reason for this observation is the relatively modest number of meioses that separated most of the thousand pairs that were used in our comparison. With a lower number of separating meioses, in general, not many mutations will have accumulated. In the case of RM Y-STRs, which also includes many multi-copy loci, however, some relatively closely related pairs may display multiple mutational steps in a multi-copy locus. SMM and MLM consider those as individual mutations, while IAM only considers two states: mutated or not-mutated. In principle this could explain the larger degree of overestimations as observed with SMM and MLM (Figure 6d). In addition, the assumption that multi-step variations between pairs were the result of the result of multiple single-step mutations rather than a single multi-step mutation may have had an impact on rate of overestimations observed in SMM and MLM. More comprehensive future studies may shed more light on the differences that are observed between various studies, Y-STR kits and models.

Patrilineal investigative genetic genealogy

Forensic genetic genealogy, also referred to as investigative genetic genealogy (IGG) has provided many success stories on solving cold cases over de recent years in different countries, especially the US [44]. The technology relies on the use of large numbers of autosomal SNPs, typically obtained with genome-wide SNP microarrays, and databases that contain such large SNP data, such as direct-to-consumer databases. These databases were created using genetic material from people that provided their DNA for other purposes than crime-solving, such as finding family members for private reasons [45]. This has evoked many ethical [46-48], but also some technical [49] concerns. Another approach is to use kinship analysis using autosomal STR profiles already included in criminal offender databases [50]. However, autosomal STRs can typically only detect first-degree relatives (i.e., parent-offspring, or siblings) with high statistical certainty.

Y-STRs also have the potential to aid in the identification of criminals and missing persons via patrilineal investigative genetic genealogy, in cases where no autosomal STR matches can be found in national forensic DNA databases [51]. Currently, only autosomal STRs are included in most national criminal offender DNA databases. Thus, the first step to implement patrilineal investigative genetic genealogy, would be to start complementing autosomal STR profiles with Y-STR profiles – preferably using new generation commercial kits, like Yfiler Plus, or PowerPlex® Y23, in national criminal offender DNA databases. Y-STR haplotypes generated by such kits have a high discrimination power for unrelated individuals, but a rather low power for differentiating patrilineally related men (as also demonstrated here for Yfiler Plus). However, this represent an advantage for patrilineal investigative genetic genealogy as we envision. Crime scene traces from male donors that did not result in a match based on autosomal STRs, simply because the trace donor was not present in the criminal offender DNA database, could demonstrate a Y-STR match with a paternal relative of the unknown trace donor that was present in the database. Such a Y-STR match based on a new generation commercial Y-STR kit would then function as a starting point for further investigations to find the paternally related unknown trace donor. Additional analysis using the RMplex in the trace sample and the reference DNA sample of the matching relative in the database would be useful to find out if the unknown trace donor is a close or a distant relative of the man in the database, which provides important investigative information to find the unknown perpetrator. Moreover, if there are several Y-STR matches in the criminal offender DNA database, RMplex will help in separating close from distant relatives and allows police investigation to focus on the close relatives in search for the unknown perpetrator.

Of course, all these applications described for Y-chromosomal familial search in forensic databases would also work in familial search based on voluntary mass screenings. That way the Marianne Vaatstra case in the Netherlands was investigated and would have been solved also in case the unknown perpetrator would not have willingly participated in the voluntary mass screening, which he did [5]. The advantage of this approach is that it could be easily integrated in routine forensic casework, as STR typing and storing the genotypes in national databases is already common practice. Applying autosomal SNP based investigative genetic genealogy, on the other hand, is a lot more divergent from the classical forensic genetic casework. As standard Y-STR profiles typically will be shared by several, or even over a dozen of males, this approach would amplify the reach of normal forensic databases several folds. Just as with autosomal SNP based forensic genetic genealogy, patrilineal investigative genetic genealogy could ultimately lead to cold cases being solved, missing persons being identified and crimes being prevented as perpetrators could be caught before committing their next crimes. The empirical evidence presented in the current study shows that by combining the strengths of standard Y-STR kits and RMplex, the patrilineal forensic genetic genealogy approach could become effective when applied on a large scale.

Clearly, legislation that allows the storage of Y-STR profiles in national forensic offender DNA databases and its use for patrilineal familial search as well as for Y-STR based familial search in voluntary mass screenings has to come first. In the Netherlands, for instance, such legislation is in place since 2012, and Y-STR based familial search in voluntary mass screenings has been used in several high-profile cold cases since the Vaatstra case. However, despite of the success stories that do exist with voluntary Y-STR based mass screenings also in other countries [52], to our knowledge, the standard inclusion of Y-STR profiles in the criminal offender DNA database has not yet been adopted in any country, also not in the Netherlands. Although it has not escaped our notice that China has been making large steps in that direction for years and thousands of criminal cases were already solved because of it [51, 53]. The generation of Y-STR profile from reference samples of every offender in the national forensic DNA database will need time before this approach can show its effectiveness in forensic practice such as cold cases with unknown male perpetrators. Moreover, the societal impact would need to be carefully considered, as innocent men who are merely related to a criminal offender will become the subjects of police investigations. Discussions regarding ethics, genetic privacy, proportionality, etc. for including Y-STR profiles in criminal offender DNA databases to allow patrilineal familial searches go well beyond the scope of this study. However, this study does show that from a scientific and technological perspective there is a great

potential to increase crime-solving rates using Y-STRs, in particular by the added value provided by Y-STRs with increased mutation rates such as the 30 markers included in RMplex.

Conclusions

The study presented here shows that using pedigrees is an efficient approach to obtain empirical estimates of mutation rates and male relative differentiation rates for Y-STRs, including Y-STRs with increased mutation rates as studied here. We demonstrated that with RMplex a large proportion of closely and nearly all of distantly related males of different degrees of relationship can be differentiated, while much lower differentiation rates are achieved with the state-of-the-art commercial Y-STR kit Yfiler Plus. We show that predicting the degree of patrilineal consanguinity based on Y-STR data is feasible and that Y-STRs with high mutation rates such as those in RMplex delivered more precise prediction results than Y-STRs with lower mutation rates such as those in Yfiler Plus. Lastly, we emphasize that implementing new strategies involving Y-STRs with lower mutability and others with high mutability in routine forensic practice will open up new avenues to solve crimes that would otherwise remain unsolved.

Materials & methods

DNA samples

Within this study, a total of 1,793 male DNA samples were analyzed. These males belonged to a total of 403 pedigrees from three cohorts. Cohort 1 consisted of a total of 1,075 Dutch males belonging to 201 male pedigrees ; in total Cohort 1 spanned 1856 meioses. The samples included in Cohort 1 were collected in the context of the Erasmus Rucphen Family study [54]. The Erasmus Rucphen Family study protocol was approved by the Medical Ethics Committee of the Erasmus MC Rotterdam, the Netherlands (MEC 213.575/2002/114). In accordance with the Declaration of Helsinki, the Erasmus Rucphen Family study obtained informed consent from all participants prior to their entering the study. Cohort 2 consisted of a total of 265 males belonging to 105 male pedigrees. All males in this cohort had either the Dutch or the Belgian nationality (the Belgian males all came from the Flemish part of Belgium); in total Cohort 2 spanned 2,089 meioses. The Medical Ethics committee at KU Leuven/UZ Leuven allowed broad Y-STR analyses of the patrilineal relatives (S55864; S59085; S54010) from Cohort 2. The larger cohort to which these samples belonged are described in more detail elsewhere [9]. Cohort 3 consisted of 453 males belonging to 97 pedigrees. All males in this cohort had the Pakistani nationality

and had been part of a previous study into RM Y-STRs [19]; in total Cohort 3 spanned 405 meioses. The Ethics board from University of Health Sciences Lahore Pakistan approved the collection of the samples from Cohort 3 (UHS/Education/126-13/2390), which were all collected under informed consent. All participants of all three cohort studies provided informed consent. The Medical Ethics Committee of the Erasmus MC allowed the execution of the present study within the Erasmus MC (MEC-2020-0535).

The different cohorts have different characteristics, where Cohort 2 consist mostly of males that share distant common paternal ancestors, Cohort 3 is characterized by containing closely related males. Cohort 1 contains pedigrees with large numbers of males with both recent and more distant common paternal ancestors, albeit not as distant as could be found in Cohort 2. Figure S3 visualizes the differences between the different cohorts with regard to the total number of male relative pairs and the degree of consanguinity between those pairs. Table 1 provides summary statistics that show the difference between the three cohorts.

Table 1: Summary statistics of the three cohorts included in this study.

	Cohort 1	Cohort 2	Cohort 3
Individuals	1075	265	453
Number of pedigrees	201	105	97
Mean number of individuals per pedigree	5.4	2.5	4.7
Median number of individuals per pedigree	2	2	4
Max number of individuals per pedigree	50	16	10
Total meiosis covered	1856	2089	405
Mean number of meioses between pairs	7.86	17.51	2.29
Median number of meioses between pairs	8	17	2
Biogeographic ancestry	Northwestern Europe	Northwestern Europe	South Asian

Y-STR Genotyping

All males were genotyped using RMplex for 30 Y-STRs with increased mutation rates under the conditions as described previously [24], using the alternative primer for DYS570 and

reducing the total reaction volume to 10 µL. Additionally, the males from Cohort 1 were also typed using Yfiler™ Plus PCR Amplification Kit (Thermo Fisher Scientific) following the manufacturer's protocols, except for a reduced total reaction volume of 10 µL. All amplifications were performed on a Veriti™ 96-Well Fast Thermal Cycler (Thermo Fisher Scientific). Capillary electrophoreses were performed on a 3500 Series Genetic Analyzer (Thermo Fisher Scientific) equipped with a 36 cm 8-capillary array and using POP-4 (Thermo Fisher Scientific). GeneScan™ 600 LIZ™ dye Size Standard v2.0 (Thermo Fisher Scientific) was used as internal size standard. The interpretation of the electropherograms was performed using GeneMapper® ID-X Software Version 1.5 (Thermo Fisher Scientific).

Estimating mutation rates from pedigrees

To estimate the mutation rates using the pedigree information, we used the frequentist approach where the mutation rate was defined as the total number of observed mutations divided by the total number of meioses. This analysis was performed for each pedigree, the numbers of mutations and meioses from each pedigree were summed per cohort, and lastly the per-marker mutation rates were estimated by combining the three cohorts together. Clopper-Pearson intervals were used to indicate the uncertainty of the mutation rate estimates.

When estimating the number of mutations based on pedigree data (instead of father-son pairs) there is a need to make certain assumptions, as pedigrees may include males separated by many generations while the analyzed males only come from the more recent generations. The first assumption that was made, was that if no haplotypic difference was observed between a pair of males connected by individuals of which no data was available, that no mutation had occurred among all these males. The second assumption was that if multistep mutations were observed between two patrilineally related males, that this should be explained as multiple single step mutations rather than a single multistep mutation. The exception to the latter was in cases where the multi-step variation were found in a father-son pair, since in such cases a single multistep mutation was the only valid explanation. Furthermore, our approach always assumed the lowest number of mutations to explain the genotype variability between the individuals within a pedigree. These assumptions are expected to hold true in the majority of cases, but may lead to errors in some cases. Figure S4 shows an example of how the number of mutations were estimated in this study. In this example, a total of five mutations were concluded. Individual A-F shared the same mutation, which was most likely inherited from their most recent common ancestor; hence, these variations could be explained by a single mutation. Alternatively, the genotypes could be explained by three parallel mutations; however,

observing the same mutations three times in three brothers (A, B, and C) independently is highly unlikely and therefore this scenario was rejected. The same mutation was also observed in individual N; as this mutation is not shared by any of the close relatives of this individual, the most probable explanation is an independent mutation that took place in this individual. The other variation that was observed in this example pedigree was a mutation from allele 10 to allele 8 which was observed in two individuals. In individual T it could only be explained by a single two-step mutation, as there was also data available from the father of individual T (i.e., individual Q), where the mutation was not present. In individual T, however, there was no data available from the father or any other close paternal relative. Hence, for this individual it was assumed that two single-step mutations would be the most probable explanation; these mutations could have taken place at individual U, or at any of his three direct paternal ancestors. Importantly, the possibility that, just as in individual T, a single two-step mutation had taken place in one of these individuals cannot be ruled out based on the available data.

The most simple scenarios are encountered when dealing with single-copy Y-STRs, for example if one individual has allele 10 for a given Y-STR while a second individual from the same pedigree carries allele 12 for that same Y-STR, it will be assumed that two mutations had occurred. In contrast, multi-copy loci can lead to more complex scenarios; for example, in Figure S5a the most straightforward solution (and the one that was assumed) is if allele 10 from individual A had mutated to allele 9 in individual B, so only one mutation had occurred. Alternatively, allele 10 from individual A could have mutated to allele 11 in individual B, while allele 11 in individual A mutated to allele 9 in individual B, this would require three mutational steps; although less likely such a scenario would not be impossible. Figure S5b shows a scenario where individual B carries a microvariant allele, while individual A does not. Here we considered the step from a microvariant allele to an adjacent conventional allele as one mutational step; hence, in Figure S5b the mutation from allele 10 to 9.2 is considered as one mutation. In general, the scenario with the lowest number of mutation steps is preferred, Figureure S5c, however, shows an exception. If two individuals carried a microvariant allele, is was assumed that those two alleles are derived from the same copy; therefore in a situation as encountered in Figureure S5c, we would consider allele 11.2 to have mutated to 9.2 and allele 10 to 11, although mutations from allele 10 to 9.2 and from allele 11.2 to 11, respectively, would have explained the genotypes with less mutational steps. Lastly, Figureure S5d shows an example where two individuals have a different number of detected alleles in a multi-copy Y-STRs. For the genotyping we did not take peak heights into account for reasons explained elsewhere [24], meaning that even if in individual B, allele 11 would show twice the height of allele 14, we would still call the genotype as 11, 14, instead of 11, 11, 14. In

such cases too, the path with the lowest mutation steps is assumed, in this example that means that allele 10 in individual A would have likely mutated to allele 11 in individual B, ergo individual B would carry two copies of allele 11.

Estimating differentiation rates

The frequentist approach was also used to calculate the male relative differentiation rates for every group of relatives separated by one to 34 meioses. Here, pairwise comparisons of all individuals within each pedigree were made, to identify all pairs of relatives that were separated by a certain number of meioses. From each pair separated by a given number of meioses the number of observed mutations between the individuals within the pair was assessed. The differentiation rate for given number of separating meioses (i.e., one to 34 in the total dataset) was calculated by dividing the number of pairs that displayed at least one allelic difference at one Y-STR marker, by the total number of pairs with that number of separating meioses. A comparative analysis between Yfiler Plus and RMplex was done on individuals from Cohort 1, as the sample size and the structure of the pedigrees in this cohort allowed to make a comprehensive assessment of the differentiation rate in a range of one to thirteen meioses. Clopper-Pearson intervals were estimated to indicate the statistical uncertainties of the differentiation rate estimates.

Prediction of the degree of patrilineal consanguinity using a machine leaning based model (MLM)

A machine learning approach was used to attempt to predict the number of meioses that separated a pair of relatives based on the observed Y-STR genotype differences. In order to train the models, data were simulated based on the reference mutation rate estimates for all Y-STRs derived from a recent study that combined data from many father-son based studies [25]. For each number of separating meioses in the range of one to fifty, a total of 100,000 pairs were simulated (5 million data points in total per model). The probability of a mutation occurring at each individual Y-STR was set to be equal to the mutation rate. Once a mutation was simulated for a given Y-STR, the probability that it would mutate further in the next generation was half of the mutation rate, as was the probability that it would mutate back to the base position (i.e., no observed allelic difference between the pair for the given Y-STR). Moreover, the probability of a single two-step mutation occurring was set 3% of the total mutation probability. For multi-copy Y-STRs, each copy

was simulated independently where the probability of a mutation occurring was equal to the mutation rate divided by the number of copies.

The simulated dataset was used to train models; the model used was a multilayer perceptron classifier as implemented by the python package scikit-learn [55]. We classified between one and 50 separating meioses based on a number of pre-determined sets of Y-STRs (Yfiler Plus, RMplex, and both assays combined). The model was trained using the default of 1 input layer, one hidden layer, one output layer, and otherwise, the default parameters for scikit-learns multilayer perceptron were used. The function randomizedSearchCV was used to randomly select the learning_rate, activation, alpha beta_1, beta_2, and the number of nodes in the hidden layer from a pre-defined feature space. In total 1,000 different combinations of parameters where tested and each validated with a two-fold cross validation step using the StratifiedKFold function of scikit-learn [55].

The resulting models were validated using the empirical data generated in the context of this study. For each pair, and for each of the three Y-STR assays, the model assigned probabilities to each category, ranging from one to fifty separating meioses. Using those probabilities, prediction intervals were calculated at 85%, 95%, and 99% probability. These prediction intervals were determined by summing up the probabilities obtained for each of the individual meiotic distances. To find the optimal prediction interval multiple cycles of testing were performed, the size of the window was increased each cycle and then slid through all the possible combinations of adjacent meiotic distances. Once the predefined confidence level was reached using this approach the narrowest prediction interval that resulted in the largest combined probability was returned as the prediction interval. The prediction accuracy of the models was determined by calculating the proportions of relative pairs where the true number of separating meioses fell within the respective predicted intervals. Additionally, to evaluate the precision of the different models, the size of the intervals was evaluated amongst the different assays and different number of separating meioses.

Comparison with different prediction models

To compare the newly developed machine learning based models with established models as described by Walsh [56], the R-script developed by Boattini *et al.* was implemented [30]. A random sub selection of a thousand pairs from the three cohorts was made (the distribution of different relationships is shown in Figure S3). The number of mutational steps for those pairs were derived and used as input for SSM and MLM. The data had to

be slightly modified where all non-zero values were transformed to the value 1 to serve as input for IAM. The R-script for IAM could be applied unmodified; however, SMM required a small modification as the high mutation rates found in RMplex led to errors. The numbers became bigger than the maximum floating point number in R of approximately 1.8e308. To overcome this error the "Rmpfr" packages (https://CRAN.R-project.org/package=Rmpfr) was used to allow for calculations up to 128 bit floating point numbers. The average mutation rate was derived from the same reference as used previously [25] to match the mutation rates as used by MLM. The resulting 95% confidence intervals described the number of meioses to the common ancestor. Since MLM rather predicts the number of meioses separating the pair the intervals obtained from IAM and SSM were multiplied by a factor two. The lower point was rounded down and the upper bound was rounded up as the true number of separating meioses is always an integer.

Data visualization

Plots of pedigree structures were made using yEd (https://www.yworks.com/products/yed). Graphs were made using Rstudio in combination with the "ggplot2" packages [57]. Venn diagram were made in Rstudio using the "ggven" packages. The probability graphs in Figure S1 were made using the online tool presented in this publication which can be found on ystr.erasmusmc.nl.

Acknowledgements

The authors are grateful to all participants of all cohort studies. We thank Cornelia van Duijn and Ben Oostra for setting-up the Erasmus Rucphen Family (ERF) study as well as P. Veraart for help with sorting out the genealogy records, J. Vergeer and P. Snijders for their help in retrieving the materials needed to analyze Cohort 1. We additionally thank Jan Geypen for sample collection and follow-up for Cohort 2. Ronny Decorte is acknowledged for useful comments on the manuscript.

References

1. Roewer, L., et al., *Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts.* Human Genetics, 1992. **89**(4): p. 389-394.
2. Roewer, L. and J.T. Epplen, *Rapid and sensitive typing of forensic stains by PCR amplification of polymorphic simple repeat sequences in case work.* Forensic science international, 1992. **53**(2): p. 163-171.
3. Kayser, M., et al., *Evaluation of Y-chromosomal STRs: a multicenter study.* International journal of legal medicine, 1997. **110**(3): p. 125-133.
4. Prinz, M., et al., *Multiplexing of Y chromosome specific STRs and performance for mixed samples.* Forensic Science International, 1997. **85**(3): p. 209-218.
5. Kayser, M., *Forensic use of Y-chromosome DNA: a general overview.* Human Genetics, 2017. **136**(5): p. 621-635.
6. Kayser, M. and A. Sajantila, *Mutations at Y-STR loci: implications for paternity testing and forensic analysis.* Forensic Science International, 2001. **118**(2-3): p. 116-121.
7. Brinkmann, B., et al., *Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat.* American Journal of Human Genetics, 1998. **62**(6): p. 1408-1415.
8. Calafell, F. and M.H.D. Larmuseau, *The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research.* Human genetics, 2017. **136**(5): p. 559-573.
9. Larmuseau, M.H.D., et al., *A historical-genetic reconstruction of human extra-pair paternity.* Current biology, 2019. **29**(23): p. 4102-4107. e7.
10. King, T.E. and M.A. Jobling, *What's in a name? Y chromosomes, surnames and the genetic genealogy revolution.* Trends in genetics, 2009. **25**(8): p. 351-360.
11. Xu, H., et al., *Inferring population structure and demographic history using Y-STR data from worldwide populations.* Molecular genetics and genomics, 2015. **290**(1): p. 141-150.
12. Cai, X., et al., *Human migration through bottlenecks from Southeast Asia into East Asia during Last Glacial Maximum revealed by Y chromosomes.* PloS one, 2011. **6**(8): p. e24282.
13. Myres, N.M., et al., *A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe.* European Journal of Human Genetics, 2011. **19**(1): p. 95-101.
14. Ballantyne, K.N. and M. Kayser, *Additional Y-STRs in Forensics: Why, Which, and When.* Forensic Science Review, 2012. **24**(1): p. 63-78.

15. Larmuseau, M.H.D., et al., *Recent radiation within Y-chromosomal haplogroup R-M269 resulted in high Y-STR haplotype resemblance.* Annals of human genetics, 2014. **78**(2): p. 92-103.
16. de Knijff, P., *On the Forensic Use of Y-Chromosome Polymorphisms.* Genes, 2022. **13**(5): p. 898.
17. Ballantyne, K.N., et al., *Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications.* American Journal of Human Genetics, 2010. **87**(3): p. 341-353.
18. Ballantyne, K.N., et al., *A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages.* Forensic Science International: Genetics, 2012. **6**(2): p. 208-218.
19. Adnan, A., et al., *Improving empirical evidence on differentiating closely related men with RM Y-STRs: a comprehensive pedigree study from Pakistan.* Forensic Science International: Genetics, 2016. **25**: p. 45-51.
20. Ballantyne, K.N., et al., *Toward Male Individualization with Rapidly Mutating Y-Chromosomal Short Tandem Repeats.* Human Mutation, 2014. **35**(8): p. 1021-1032.
21. Gopinath, S., et al., *Developmental validation of the Yfiler® Plus PCR Amplification Kit: An enhanced Y-STR multiplex for casework and database applications.* Forensic Science International: Genetics, 2016. **24**: p. 164-175.
22. Thompson, J.M., et al., *Developmental validation of the PowerPlex® Y23 System: a single multiplex Y-STR analysis system for casework and database samples.* Forensic Science International: Genetics, 2013. **7**(2): p. 240-250.
23. Ralf, A., et al., *Identification and characterization of novel rapidly mutating Y-chromosomal short tandem repeat markers.* Human Mutation, 2020. **41**(9): p. 1680-1696.
24. Ralf, A., et al., *RMplex: An efficient method for analyzing 30 Y-STRs with high mutation rates.* Forensic Science International: Genetics, 2021(55): p. 102595.
25. Neuhuber, F., et al., *Improving the differentiation of closely related males by RMplex analysis of 30 Y-STRs with high mutation rates.* Forensic Science International: Genetics, 2022: p. 102682.
26. Burgarella, C. and M. Navascués, *Mutation rate estimates for 110 Y-chromosome STRs combining population and father–son pair data.* European Journal of Human Genetics, 2011. **19**(1): p. 70.
27. Yuan, L., et al., *Mutation analysis of 13 RM Y-STR loci in Han population from Beijing of China.* International Journal of Legal Medicine, 2019. **133**(1): p. 59-63.

28. Zhang, W., et al., *Multiplex assay development and mutation rate analysis for 13 RM Y-STRs in Chinese Han population.* International Journal of Legal Medicine, 2017. **131**(2): p. 345-350.
29. Boattini, A., et al., *Mutation rates and discriminating power for 13 rapidly-mutating Y-STRs between related and unrelated individuals.* PLOS One, 2016. **11**(11): p. e0165678.
30. Boattini, A., et al., *Estimating Y-Str Mutation Rates and Tmrca Through Deep-Rooting Italian Pedigrees.* Scientific Reports, 2019. **9**(1): p. 9032.
31. Claerhout, S., et al., *Determining Y-STR mutation rates in deep-routing genealogies: Identification of haplogroup differences.* Forensic Science International: Genetics, 2018. **34**: p. 1-10.
32. Larmuseau, M.H.D., K. Matthijs, and T. Wenseleers, *Cuckolded fathers rare in human populations.* Trends in ecology & evolution, 2016. **31**(5): p. 327-329.
33. Claerhout, S., et al., *A game of hide and seq: Identification of parallel Y-STR evolution in deep-rooting pedigrees.* European Journal of Human Genetics, 2019. **27**(4): p. 637.
34. Claerhout, S., et al., *Determining Y-STR mutation rates in deep-routing genealogies: identification of haplogroup differences.* Forensic Science International: Genetics, 2018.
35. Ambrosio, I.B., et al., *Mutational data and population profiling of 23 Y-STRs in three Brazilian populations.* Forensic Science International: Genetics, 2020. **48**: p. 102348.
36. Javed, F., et al., *Male individualization using 12 rapidly mutating Y-STRs in Araein ethnic group and shared paternal lineage of Pakistani population.* International Journal of Legal Medicine, 2018. **132**(6): p. 1621-1624.
37. Zgonjanin, D., et al., *Mutation rate at 13 rapidly mutating Y-STR loci in the population of Serbia.* Forensic Science International: Genetics Supplement Series, 2017. **6**: p. e377-e379.
38. Roewer, L., et al., *DNA commission of the International Society of Forensic Genetics (ISFG): Recommendations on the interpretation of Y-STR results in forensic analysis.* Forensic Science International: Genetics, 2020. **48**: p. 102308.
39. Roewer, L., et al., *DNA commission of the International Society of Forensic Genetics (ISFG): Recommendations on the interpretation of Y-STR results in forensic analysis*. 2020, Elsevier. p. 102308.
40. Roewer, L., *Y-chromosome short tandem repeats in forensics—Sexing, profiling, and matching male DNA.* Wiley Interdisciplinary Reviews: Forensic Science, 2019. **1**(4): p. e1336.
41. Andersen, M.M. and D.J. Balding, *How convincing is a matching Y-chromosome profile?* PLOS Genetics, 2017. **13**(11): p. e1007028.

42. Claerhout, S., et al., *CSYseq: The first Y-chromosome sequencing tool typing a large number of Y-SNPs and Y-STRs to unravel worldwide human population genetics.* PLOS Genetics, 2021. **17**(9): p. e1009758.
43. Claerhout, S., et al., *YMrCA: Improving Y-chromosomal ancestor time estimation for DNA kinship research.* Human mutation, 2021. **42**(10): p. 1307-1320.
44. Dowdeswell, T.L., *Forensic genetic genealogy: A profile of cases solved.* Forensic Science International: Genetics, 2022. **58**: p. 102679.
45. Kling, D., et al., *Investigative genetic genealogy: Current methods, knowledge and practice.* Forensic Science International: Genetics, 2021. **52**: p. 102474.
46. Berkman, B.E., W.K. Miller, and C. Grady, *Is it ethical to use genealogy data to solve crimes?* Annals of Internal Medicine, 2018. **169**(5): p. 333-334.
47. Samuel, G. and D. Kennett, *The impact of investigative genetic genealogy: perceptions of UK professional and public stakeholders.* Forensic Science International: Genetics, 2020. **48**: p. 102366.
48. Guerrini, C.J., et al., *Four misconceptions about investigative genetic genealogy.* Journal of Law and the Biosciences, 2021. **8**(1): p. lsab001.
49. de Vries, J.H., et al., *Impact of SNP microarray analysis of compromised DNA on kinship classification success in the context of investigative genetic genealogy.* Forensic Science International: Genetics, 2022. **56**: p. 102625.
50. Ge, J. and B. Budowle, *How many familial relationship testing results could be wrong?* PLoS genetics, 2020. **16**(8): p. e1008929.
51. Ge, J. and B. Budowle, *Forensic investigation approaches of searching relatives in DNA databases.* Journal of Forensic Sciences, 2021. **66**(2): p. 430-443.
52. Dettlaff-Kakol, A. and R. Pawlowski, *First Polish DNA" manhunt"–an application of Y-chromosome STRs.* International Journal of Legal Medicine, 2002. **116**(5): p. 289-291.
53. Ge, J., et al., *Future directions of forensic DNA databases.* Croatian medical journal, 2014. **55**(2): p. 163.
54. Sayed-Tabatabaei, F.A., et al., *Heritability of the function and structure of the arterial wall: findings of the Erasmus Rucphen Family (ERF) study.* Stroke, 2005. **36**(11): p. 2351-2356.
55. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* Journal of machine Learning research, 2011. **12**: p. 2825-2830.
56. Walsh, B., *Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals.* Genetics, 2001. **158**(2): p. 897-912.

57. Wickham, H., *ggplot2.* Wiley interdisciplinary reviews: Computational statistics, 2011. **3**(2): p. 180-185.

Supporting information

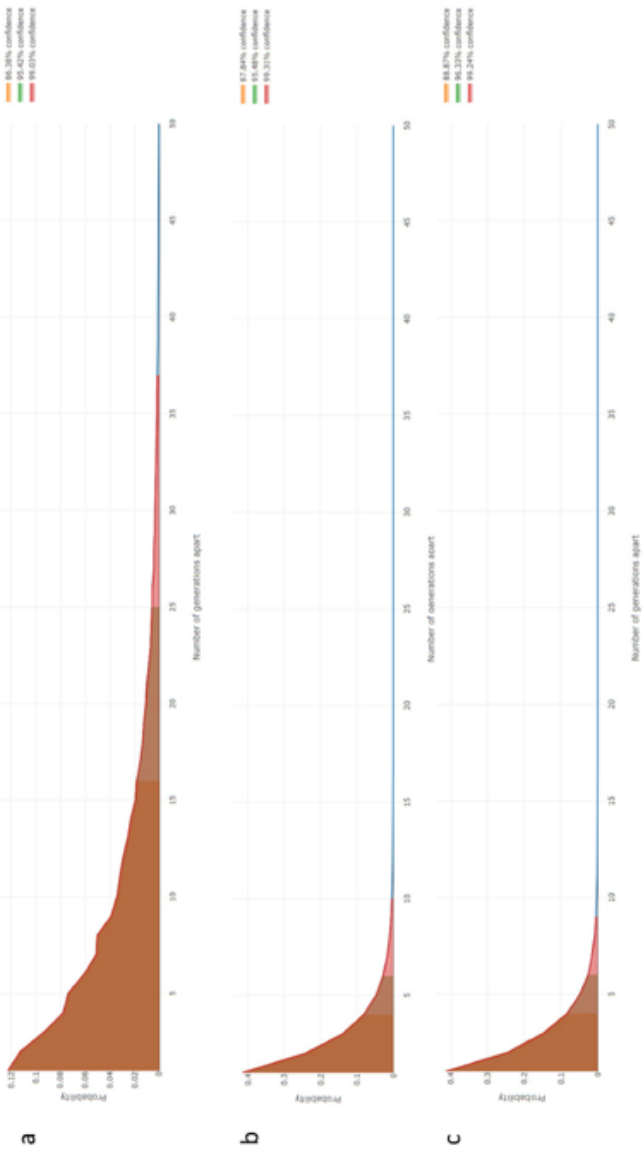


Figure S1: Predicted intervals with at least 85%, 95% and 99% probability with Yfiler Plus (a), RMplex (b) and the two assays combined (c) when no mutations between two individuals are observed.

.

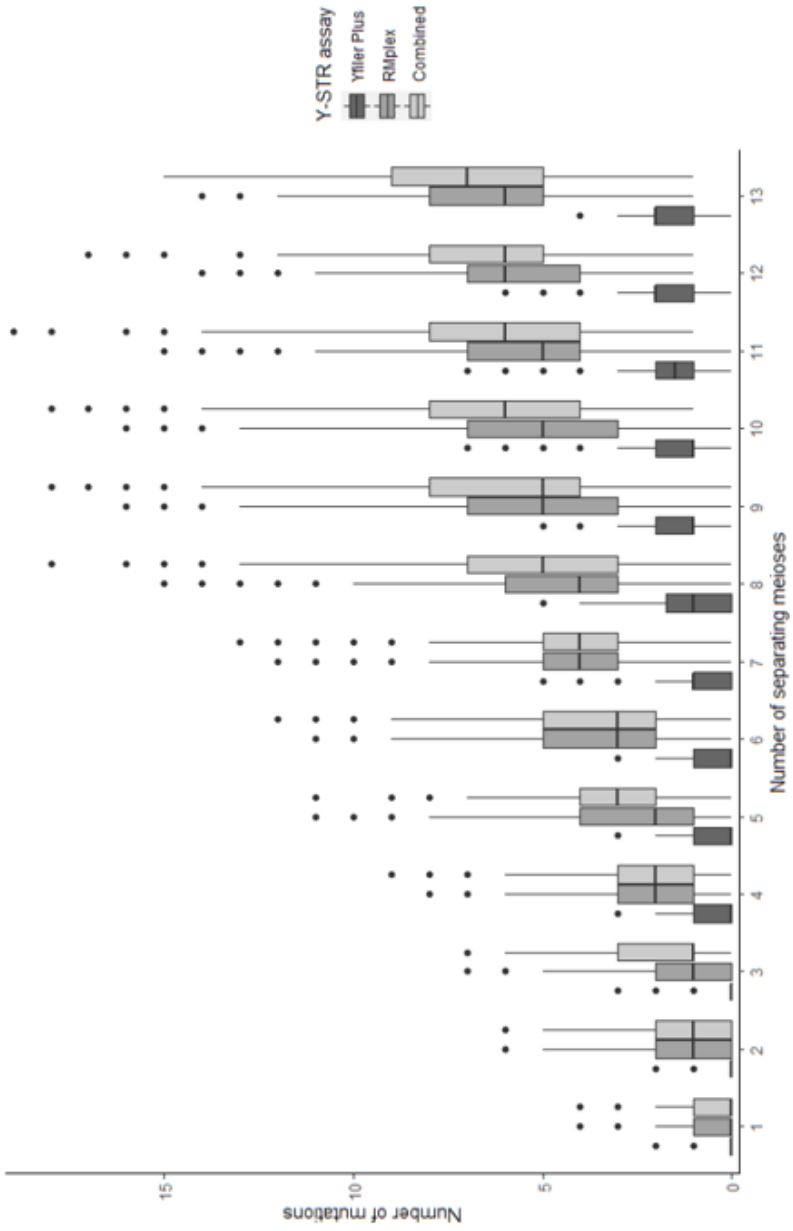


Figure S2: Boxplots showing the accumulation of mutations with increasing numbers of meioses.

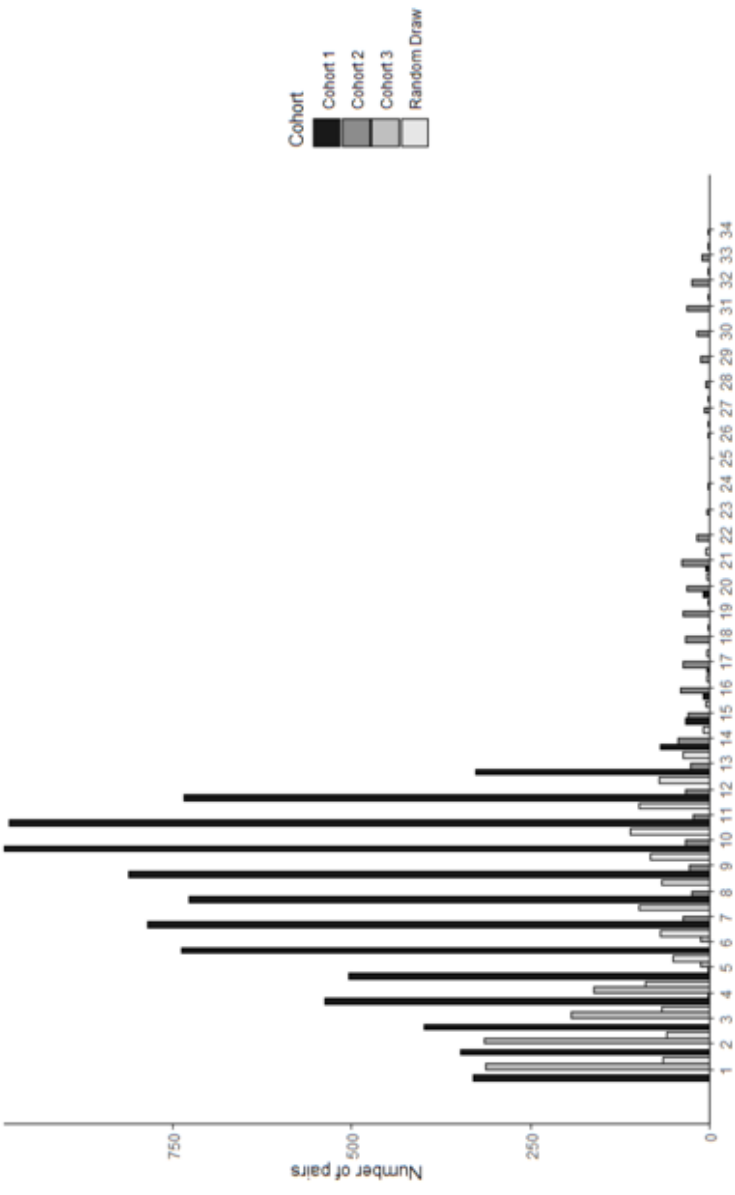


Figure S3: Distribution of number of separating meioses for all pairs of males in the different cohorts and for the 1000 randomly drawn pairs for the prediction model comparisons.

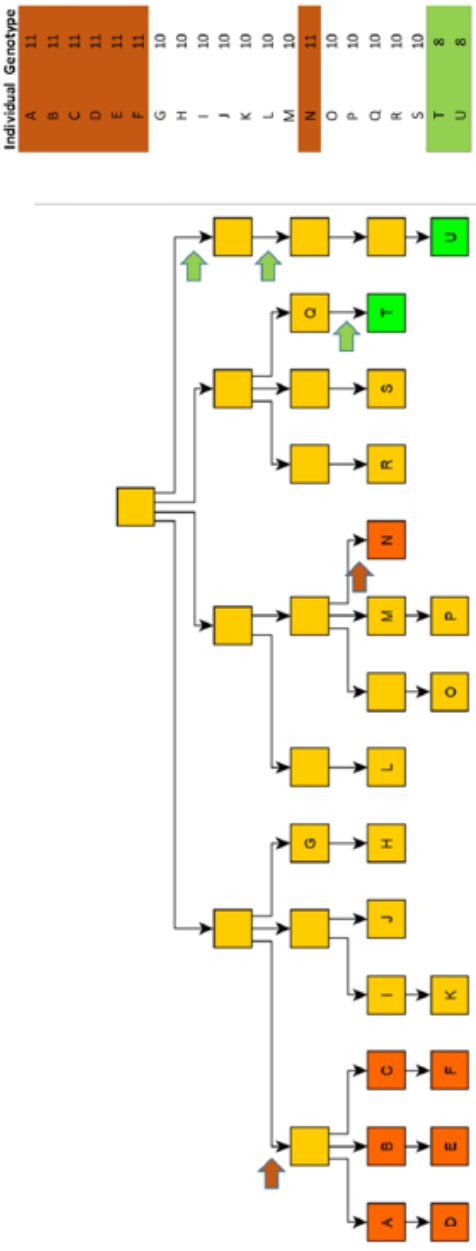


Figure S4: Example of a pedigree spanning 35 meioses, where a total of five mutations were estimated based on the observed genotype variations.

a.

Indv. A \ Indv. B	9	11	16
10	-1	+1	+6
11	-2	0	+5
16	-7	-5	0

b.

Indv. A \ Indv. B	9.2	11	16
10	-1	+1	+6
12	-3	-1	+4
16	-7	-5	0

c.

Indv. A \ Indv. B	9.2	11	16
10	-1	+1	+6
11.2	-2	-1	+5
16	-7	-5	0

d.

Indv. A \ Indv. B	11	14
10	+1	+4
11	0	+3
16	-5	-2

Figure S5: Various scenarios that could be encountered when dealing with multi-copy loci, the solutions that our approach preferred in these cases are highlighted.

Table S1: Combined mutation rate estimates as derived from all pedigreed included within the three analyzed cohorts.

Marker	Total meioses	Mutations	Mutation rate ($x10^{-3}$)	95% confidence range ($x10^{-3}$)	Reference value [1] ($x10^{-3}$)	Fisher's exact p-value
DYF399S1	4354	305	70.1	62.6 - 78.0	62.8	0.125
DYF1001	4354	233	53.5	47.0 - 60.6	48.0	0.372
DYF1000	4354	211	48.5	42.3 - 55.3	35.9	**0.018**
DYF403S1a	4354	168	38.6	33.1 - 44.7	27.3	**0.001**
DYS724	4354	167	38.4	32.8 - 44.5	48.0	0.074
DYS711	4354	131	30.1	25.2 - 35.6	26.6	0.479
DYS712	4354	113	26.0	21.4 - 31.1	31.1	0.221
DYS612	4354	99	22.7	18.5 - 27.6	16.3	**0.012**
DYS1012	4354	97	22.3	18.1 - 27.1	15.8	0.091
DYR88	4354	94	21.6	17.5 - 26.4	26.3	0.253
DYS1010	4354	84	19.3	15.4 - 23.8	14.0	0.133
DYF1002	4354	78	17.9	14.2 - 22.3	16.8	0.841
DYS518	4354	71	16.3	12.8 - 20.5	13.3	0.152
DYF404S1	4354	61	14.0	10.7 - 18.0	12.5	0.509
DYS547	4354	60	13.8	10.5 - 17.7	14.7	0.751
DYS627	4354	59	13.6	10.3 - 17.4	14.5	0.709
DYS1007	4354	59	13.6	10.3 - 17.4	17.2	0.274
DYS526b	4354	57	13.1	9.9 - 16.9	12.3	0.735
DYS713	4354	53	12.2	9.1 - 15.9	13.9	0.555
DYS626	4354	50	11.5	8.5 - 15.1	8.6	0.122
DYS449	4354	45	10.3	7.5 - 13.8	11.2	0.673
DYS570	4354	44	10.1	7.4 - 13.5	8.3	0.295
DYF393S1	4354	42	9.6	7.0 - 13	7.1	0.332
DYS1003	4354	41	9.4	6.8 - 12.8	12.6	0.245
DYS576	4354	40	9.2	6.6 - 12.5	12.7	0.072
DYF387S1	4354	37	8.5	6.0 - 11.7	10.2	0.363
DYS458	1860	15	8.1	4.5 - 13.3	8.5	1.000
DYF403S1b	4354	30	6.9	4.7 - 9.8	9.1	0.243
DYS1005	4354	26	6.0	3.9 - 8.7	9.8	0.118
DYS460	1860	11	5.9	3.0 - 10.6	4.3	0.343
DYS1013	4354	21	4.8	3.0 - 7.4	10.8	**0.009**
DYS385	1860	8	4.3	1.9 - 8.5	7.5	0.137
DYS390	1860	8	4.3	1.9 - 8.5	2.7	0.245
DYS391	1860	7	3.8	1.5 - 7.7	2.5	0.328
DYS389II	1860	7	3.8	1.5 - 7.7	5.5	0.487
DYS439	1860	7	3.8	1.5 - 7.7	4.8	0.713
DYS442	4354	15	3.4	1.9 - 5.7	7.4	**0.047**
DYS448	1860	6	3.2	1.2 - 7.0	0.8	**0.016**
DYS533	1860	6	3.2	1.2 - 7.0	3.5	1.000
DYS635	1860	5	2.7	0.9 - 6.3	3.8	0.541
YGATAH4	1860	4	2.2	0.6 - 5.5	1.9	0.777
DYS389I	1860	4	2.2	0.6 - 5.5	2.4	1.000
DYS456	1860	3	1.6	0.3 - 4.7	4.4	0.108
DYS481	1860	3	1.6	0.3 - 4.7	4.7	0.070
DYS19	1860	3	1.6	0.3 - 4.7	2.0	1.000
DYS438	1860	2	1.1	0.1 - 3.9	0.3	0.142
DYS393	1860	1	0.5	0.0 - 3.0	1.7	0.346
DYS437	1860	1	0.5	0.0 - 3.0	1.2	0.710
DYS392	1860	0	0.0	0.0 - 2.0	0.8	0.621

Table S2: Observed differentiation rates using RMplex in three different cohorts and all three cohorts combined.

	Cohort 1			Cohort 2			Cohort 3			Overall		
Meioses	Number of pairs	Differentiated (%)	Mean mutations (SD)	Number of pairs	Differentiated (%)	Mean mutations (SD)	Number of pairs	Differentiated (%)	Mean mutations (SD)	Number of pairs	Differentiated (%)	Mean mutations (SD)
1	324	144 (44.4)	0.6 (0.8)	0			313	132 (42.2)	0.6 (1.0)	637	276 (43.3)	0.6 (0.9)
2	332	227 (68.4)	1.2 (1.2)	0			315	200 (63.5)	1.1 (1.3)	647	427 (66.0)	1.2 (1.2)
3	385	287 (74.5)	1.6 (1.5)	0			193	154 (79.8)	1.5 (1.2)	578	441 (76.3)	1.6 (1.4)
4	524	431 (82.3)	1.9 (1.5)	1	1 (100)	2 (0)	161	141 (87.6)	1.8 (1.2)	686	573 (83.5)	2.1 (2.3)
5	489	456 (93.3)	2.8 (1.8)	1	1 (100)	7 (0)	13	13 (100)	1.6 (1.1)	503	470 (93.4)	2.8 (2.1)
6	710	681 (95.9)	3.4 (2.0)	0			12	12 (100)	1.2 (0.4)	722	693 (96.0)	3.6 (2.9)
7	762	749 (98.3)	4.0 (2.2)	23	22 (95.7)	4.8 (2.1)				785	771 (98.2)	4.1 (2.6)
8	686	676 (98.5)	4.6 (2.5)	15	15 (100)	5.2 (3.0)				701	691 (98.6)	4.9 (3.3)
9	764	757 (99.1)	5.1 (2.8)	16	16 (100)	5.5 (3.1)				780	773 (99.1)	5.3 (3.5)
10	940	936 (99.6)	5.3 (2.6)	16	16 (100)	12.4 (17.2)				956	952 (99.6)	5.7 (4.1)
11	954	952 (99.8)	5.4 (2.4)	12	12 (100)	7.4 (3.4)				966	964 (99.8)	5.6 (3.0)
12	707	707 (100)	5.8 (2.3)	23	23 (100)	7.7 (2.2)				730	730 (100)	6.0 (2.8)
13	316	316 (100)	6.3 (2.3)	18	18 (100)	8.3 (2.6)				334	334 (100)	6.6 (2.8)
14	67	67 (100)	6.6 (2.3)	16	16 (100)	8.3 (2.9)				83	83 (100)	6.9 (2.5)
15	34	34 (100)	7.9 (1.4)	12	12 (100)	7.9 (3.4)				46	46 (100)	7.9 (2.1)
16	8	8 (100)	8.6 (1.3)	25	25 (100)	10.1 (2.7)				33	33 (100)	9.7 (2.5)
17	4	4 (100)	9.3 (0.8)	23	23 (100)	10.5 (5.0)				27	27 (100)	10.3 (4.6)
18				22	22 (100)	8.8 (3.1)				22	22 (100)	8.8 (3.1)
19				20	20 (100)	8.9 (3.2)				20	20 (100)	8.9 (3.2)
20	9	9 (100)	14 (1.2)	16	16 (100)	8.1 (3.6)				25	25 (100)	10.2 (4.1)
21	5	5 (100)	14.2 (1.5)	24	24 (100)	11.4 (4.4)				29	29 (100)	11.9 (4.2)
22				10	10 (100)	14.7 (3.7)				10	10 (100)	14.7 (3.7)
23				3	3 (100)	15.7 (4.0)				3	3 (100)	15.7 (4.0)
24				1	1 (100)	8 (0)				1	1 (100)	8 (0)
25												
26				1	1 (100)	14 (0)				1	1 (100)	14 (0)
27				6	6 (100)	15.2 (3.7)				6	6 (100)	15.2 (3.7)
28				3	3 (100)	14.7 (2.6)				3	3 (100)	14.7 (2.6)
29				2	2 (100)	19.5 (4.5)				2	2 (100)	19.5 (4.5)
30				6	6 (100)	13.7 (2.8)				6	6 (100)	13.7 (2.8)
31				14	14 (100)	13.8 (2.9)				14	14 (100)	13.8 (2.9)
32				15	15 (100)	13.1 (1.6)				15	15 (100)	13.1 (1.6)
33				7	7 (100)	13.9 (1.8)				7	7 (100)	13.9 (1.8)
34				1	1 (100)	16 (0)				1	1 (100)	16 (0)

Table S3: Direct comparison of Yfiler Plus, RMplex and both assays combined on differentiation of males separated by 1-13 meioses in Cohort 1

Assay	Meioses	Pairs	Differentiated	Differentiation rate (%)	Clopper-Pearson CI lower bound	Clopper-Pearson CI upper bound
Yfiler Plus	1	324	32	9.88	6.85	13.66
Yfiler Plus	2	332	73	21.99	17.65	26.83
Yfiler Plus	3	385	88	22.86	18.76	27.38
Yfiler Plus	4	524	144	27.48	23.7	31.52
Yfiler Plus	5	489	174	35.58	31.34	40.01
Yfiler Plus	6	710	299	42.11	38.45	45.84
Yfiler Plus	7	762	391	51.31	47.7	54.92
Yfiler Plus	8	686	463	67.49	63.85	70.99
Yfiler Plus	9	764	604	79.06	76	81.89
Yfiler Plus	10	940	761	80.96	78.3	83.42
Yfiler Plus	11	954	800	83.86	81.37	86.14
Yfiler Plus	12	707	633	89.53	87.04	91.69
Yfiler Plus	13	316	284	89.87	86.01	92.97
RMplex	1	324	144	44.44	38.95	50.04
RMplex	2	332	227	68.37	63.07	73.34
RMplex	3	385	287	74.55	69.89	78.82
RMplex	4	524	431	82.25	78.71	85.43
RMplex	5	489	456	93.25	90.65	95.31
RMplex	6	710	681	95.92	94.19	97.25
RMplex	7	762	749	98.29	97.1	99.09
RMplex	8	686	676	98.54	97.34	99.3
RMplex	9	764	757	99.08	98.12	99.63
RMplex	10	940	936	99.57	98.91	99.88
RMplex	11	954	952	99.79	99.24	99.97
RMplex	12	707	707	100	99.48	100
RMplex	13	316	316	100	98.84	100
Combined	1	324	147	45.37	39.86	50.97
Combined	2	332	239	71.99	66.82	76.75
Combined	3	385	295	76.62	72.07	80.76
Combined	4	524	436	83.21	79.72	86.31
Combined	5	489	461	94.27	91.83	96.16
Combined	6	710	685	96.48	94.85	97.71
Combined	7	762	751	98.56	97.43	99.28
Combined	8	686	682	99.42	98.51	99.84
Combined	9	764	759	99.35	98.48	99.79
Combined	10	940	940	100	99.61	100
Combined	11	954	954	100	99.61	100
Combined	12	707	707	100	99.48	100
Combined	13	316	316	100	98.84	100

Chapter 7

General Discussion

In the previous chapters I described the course of my PhD project starting from the successful search for more RM Y-STRs, through the development of a method that combines these new with the already known RM Y-STRs, to finally applying this method in different types of relatives of different ethnic backgrounds. In this last chapter, I would like to reflect on the broader picture and the implications of the findings described in the earlier chapters. Moreover, I will describe how, in my vision, this specialized field of research could and should be progressed to become used in future forensic routine more widely than done today. Lastly, I will note my thoughts on how this knowledge and future research could be applied practically and more so in the future than today. Notwithstanding the beauty of fundamental science and curiosity driven research, the positive societal impact that may result from it, to me is the most rewarding aspect of the content described in this thesis.

The need for more RM Y-STRs

The effectiveness of using Y-STRs, including RM Y-STRs, in familial searching was previously described in the relation with a famous high-profile Dutch cold case [1]. However, the typing of RM Y-STRs for close male differentiation resulted in a majority of close male relatives not being differentiated (i.e., sharing the same haplotype) [2]. The need to further improve the male relative differentiation capabilities of Y-STRs was the motivation for all work described in this thesis.

Revisiting the data as described in the most extensive father-son pair based Y-STR mutation rate study [3], revealed several characteristics that could set Y-STRs with high mutation rates apart from those with low mutability. In **Chapter 2**, those characteristics were used to predict, in a large number of uncharacterized Y-STRs, which were most likely to show increased mutability. This approach led to the identification of novel RM Y-STRs as was shown by empirical data based on over 1,600 father-son pairs that were genotyped for the candidate RM Y-STRs.

The large number of 647 mutations that was observed could also be used to further confirm some of the molecular mechanisms behind Y-STR mutations. It could be shown that the long alleles mutated significantly more frequently than short alleles, which had been known from many other studies [3-9]. A less studied phenomenon that was described in **Chapter 2** was the direction of the mutation (i.e., expansion or contraction) being influenced by the allele length, where long alleles appear to be more likely to contract and short alleles expand more. This observation was not completely novel, nevertheless it independently confirmed what had been suggested in previous studies [3,

10]. Furthermore, **Chapter 2** found evidence of the age of the father at the time of conception of the child playing a role in the mutability, where older fathers on average showed more mutations compared to younger ones. This effect was previously found in some studies, while being absent, or at least not significant in others. This finding has implication for future mutation rate studies, ideally, at least the mean age of the fathers at the time of conception of the offspring should reported as it may explain a part of differences found between different studies. Lastly, **Chapter 2** described a remarkable role that the repeat motif sequence appears to play when it comes to mutability. Six out of the eight motifs that were tested show a significantly different prevalence between RM Y-STRs and non-RM Y-STRs. In particular, the motifs [AAAG], [AAGG], [AAG], and [AAG] appear to be enriched in RM Y-STRs, while other motifs appeared to be depleted from that category.

The role of these molecular characteristics of Y-STRs and their mutations were of limited importance for the following chapters of this thesis. However, they are of importance to find more RM Y-STRs in the future and also contribute towards a better understanding of what drives mutations in Y-STRs and in STRs as a whole. **Chapter 3** played a prominent part throughout this thesis as it combines the 13 previously identified RM Y-STRs [3, 11], with those newly identified in **Chapter 2** in a novel genotyping method (RMplex). In contrast to **Chapter 2**, **Chapter 3** is focusing on the technical aspect of developing and validating a CE-based genotyping assay. Although not driven by fundamental science, **Chapter 3** formed an important basis by developing the methodology that was then further applied in **Chapters 4 to 6.** Moreover, the existence of this method will allow the newly identified Y-STRs to be applied in forensic casework.

Lastly, **Chapter 4 to 6** all focused on using RMplex**,** to generate new and independent data using mainly father-son pairs (**Chapters 4 and 5**) and more extended pedigrees (**Chapter 6**)**.** In all of these chapters RMplex is compared to Yfiler™ Plus PCR Amplification Kit, the only aspect where Yfiler Plus was found to be superior to RMplex is in technical performance, as could be expected from an industry-developed genotyping kit. However, in other aspects, e.g., male relative differentiation, RMplex was far more effective than Yfiler Plus.

Considering all chapters together, the need for more RM Y-STRs was addressed and the result was a new and efficient method with unprecedented effectiveness in differentiating male relatives based on Y-STRs alone, as was shown empirically. This, however, does not mean that there is no longer a need for more RM Y-STRs. The work presented in this thesis can serve as a template for the identification and characterization of additional RM Y-STRs and RM Y-STR genotyping kits in the future.

Do rapidly mutating Y-STRs even exist?

The mutation rates of Y-STRs form a continuum ranging from as low as 10^{-5} mutations per generation (mpg) [9] until nearly 10^{-1} mpg [3]; the vast majority of Y-STRs having mutation rates in the range of 10^{-4} – 10^{-3} mpg [3, 9]. In **Chapter 2**, a four-category classification system was proposed where all Y-STRs with mutation rates below 1 x 10^{-3} mpg were classified as slowly mutating (SM) Y-STRs, the next group was termed moderately mutating (MM) Y-STRs and include Y-STRs with mutation rates in the range of 1 to 5 x 10^{-3} mpg, followed by fast mutating (FM) Y-STRs ranging from 5 x 10^{-3} to 1x 10^{-2}, every Y-STR with a mutation rate of 1 x 10^{-2} or larger was categorized as a rapidly mutating (RM) Y-STR.

However, **Chapter 4** shows the weakness of such a classification system: as an example, DYS570 had a consensus mutation rate of 1.17 x 10^{-2} mpg in European males, thus making it an RM Y-STR. In contrast, in Asian males it showed a mutation rate of 6.4 x 10^{-3} mpg and would be classified as an FM Y-STR. Overall, the mutation rate of DYS570 reached 8.3 x 10^{-3} mpg in over 11,000 males from different populations. Any classification system will have difficulties making coherent classification in the proximity of its borders. In contrast, for DYF399S1 there can be no doubt as there has not been a single study that found a mutation rate below, or even close to the 1 x 10^{-2} mpg threshold value for this Y-STR. With an overall mutation rate estimation of 6 x 10^{-2} mpg in over 7,500 father-son pairs, the RM Y-STR classification for DYF399S1 can be considered as good as irrefutable. Reverting back to examples like DYS570, it is debatable whether such Y-STRs should have been classified as RM Y-STRs in the first case, although the earliest mutation rate study on this Y-STR did suggest a mutation rate surpassing 1 x 10^{-2} mpg [3].

The difficulty to categorize Y-STRs is further accentuated in **Chapter 5,** where it is demonstrated that paternal biogeographic ancestry and the deeper evolutionary origins of Y chromosomes (i.e., Y-SNP based haplogroups) can impact the mutability of a given Y-STR. Nevertheless, the classification system as proposed in **Chapter 2** does provide a way to broadly describe the mutability of any given Y-STR, e.g., it is unlikely that an SM Y-STR in one study will be found to be an RM Y-STR in the next study. Such classifications have value when designing a specific targeted Y-STR kit. Different kits may benefit from different compositions. Some Y-STR kits may be designed for lineage identification and would thrive well by the inclusion of mostly MM Y-STRs and perhaps some RM Y-STRs. On the other hand, a Y-STR kit that is designed to uncover deep evolutionary signatures, should avoid RM Y-STRs and should rather focus on SM Y-STRs. Lastly, Y-STR kits aiming to differentiate close male relatives, as was the focus of this thesis, have little use for SM Y-STRs and should aim to include as many RM Y-STRs as possible. Until methods are developed that can genotype large numbers of Y-STRs across the entire spectrum of

mutation rates in a cost-effective manner, there will be a need to employ targeted approaches where a classification as described in Chapter 2 can help to guide the Y-STR selection.

That being said, the ambiguity of such classifications should be considered. One could also wonder, whether it practically matters whether the average locus-specific mutation rate of a given Y-STR is 9.8×10^{-3} (MM Y-STR), or rather 1.1×10^{-2} (RM Y-STR). There simply is no such thing as an unerring single mutation rate for any Y-STR locus, all we can really do is estimating based on available data and by making those estimations based on large sample sizes we can derive average mutation rate estimates that can give a reasonable approximation, yet these values cannot be simply extrapolated to other populations and certainly not to any specific male lineage.

In summary: RM Y-STRs do exist (e.g., DYF399S1), perfect locus-specific mutation rates, however, do not. Therefore, Y-STR classifications should be seen as broad indications based on limited estimations.

Can more rapidly mutating Y-STRs be identified?

The comprehensive Y-STR mutation rate study published in 2010 by Ballantyne *et al.* [3] characterized the mutation rate of all Y-STRs that had previously been identified in 2004 by Kayser *et al.* [12] based on father-son data of nearly 2000 pairs. These Y-STRs were identified in 2001 from 23 Mb Y-chromosome sequence of four genomic contigs [12]; note that this was before the completion of the Human Genome Project. In **Chapter 2** a similar approach as in Kayser *et al.* 2004 [12] was used to identify Y-STRs by using the old but still relevant software tool developed by Gray Benson in 1999: tandem repeats finder [13]. However, in **Chapter 2** the most recent genome assembly at the time (GRCh38) was used. As a result, Y-STRs that were not yet identified in 2001 and consequently were not among the 186 Y-STRs for which the mutation rates were characterized by Ballantyne *et al.* [3] could now be considered as candidate RM Y-STRs and empirically tested in a large number of father-son pairs. However, rather than genotyping all RM Y-STRs that were not included in the previous study [3], here a more focused approach was used. In Ballantyne *et al.*, less than 7% of the studied Y-STRs expressed mutation rates $>10^{-2}$ mpg [3], showing that the brute-force approach has a relatively low yield when it comes to finding RM Y-STRs. Here in **Chapter 2,** by using a more informed selection of candidate RM Y-STRs, over 44% of the studied Y-STRs were classified as RM Y-STRs. Very recently, a new genome assembly was released [14]. This assembly is thought to be the first fully completed human genome.

Obtaining this assembly, that spans previously impossible to sequence highly repetitive regions like the centromeres and telomeres, was only possible by using third-generation (i.e., long-read) sequencing technologies as have been developed by Pacific Biosciences and Oxford Nanopore Technologies [14]. It may very well be possible that doing a new scan of the new assembly would provide additional candidate Y-STRs that were not yet included in GRCh38. Moreover, some previously identified candidate RM Y-STRs were not characterized due to technical difficulties, such as: the inability to design male-specific PCR primer, poor performance in the multiplex PCR, or too challenging interpretation as the result of many stutter artifacts. Those previously excluded candidates could also be revisited and some of those technical challenges may be resolvable.

Lastly, dinucleotide repeats have been ignored in previous research, the reason is that they typically display very high stutter artifact peaks in their genotype analysis, which could interfere with the correct allele calling. However, massively parallel sequencing (MPS) may provide a solution here, not by negating the stutters produced during PCR, but by allowing to build more sophisticated stutter filters e.g. based on specific pattern recognition that can distinguish true alleles from stutter artifacts [15]. A previously developed MPS-based method already indicated that several dinucleotide Y-STRs may exhibit high mutation rates; moreover, this study demonstrated that it was possible to cope with the increased presence of stutter alleles [16].

In summary: there is a high probability that more RM Y-STRs than those identified in this thesis and before do exist, and remain to be identified. Further expanding the set of available RM Y-STRs will: 1) further enhance the capability of differentiating especially close male relatives, 2) further improve the precision of predicting the degree of patrilineal consanguinity, and will 3) further strengthen the business case for making Y-STR typing and database inclusion a standard procedure in investigative genetics.

The role of industry

Before the use of Y-STRs with high mutations rates can be more broadly implemented in forensic casework; ideally, industry should take over and develop standardized analysis methods. Especially multi-copy Y-STRs and Y-STRs frequently displaying microvariations can be more complex to interpret compared to the Y-STRs currently included in commercial genotyping assays as was shown in **Chapter 3**. Industry typically has more resources than research institutes to develop genotyping methodologies to perfection regarding critical aspect like sensitivity, specificity, reproducibility, etc. Several companies

have taken their first steps by including a limited number of Y-STRs with high mutation rates in their latest generation of Y-STR testing kits.

However, as shown in **Chapter 4-6**, the state-of-the-art Yfiler™ Plus PCR Amplification Kit, despite including six of the 13 first discovered RM Y-STRs, cannot compare to a genotyping system containing many Y-STRs with high mutation rates such as RMplex. Hitherto, companies have failed to see the relevance of developing Y-STR kits including larger numbers of Y-STRs with increased mutability. Such kits could consists of large numbers of Y-STRs with different magnitudes of mutation rates ranging from low to high for different purposes. Or, in case the multiplex capacity of available genotyping technologies is not sufficient, different Y-STR kits with sets of markers characterized by different mutation rates to be used separately (or sequentially) for different purposes e.g., a separate Y-STR kit for RM Y-STRs to be applied in cases where a match is obtained with a separate Y-STR kit for Y-STRs with lower mutation rates.

Moreover, many of the RM Y-STRs contain complex repeat structures (i.e., multiple variable repetitive stretches) making them favorable targets to massively parallel sequencing (MPS) analysis, relative to CE-based fragment length analysis used in most currently available commercial Y-STR kits. Previous research has shown that sequencing can sometimes show diversity within male lineages that would have remained undetected by fragment length analysis [17]. Targeted MPS also provides other advantages compared to CE-based analyses, i.e., there is no need to avoid overlapping allelic size ranges, which is a requirement for CE-based analyses. As a result, in principle many more Y-STRs can be targeted in a single assay, provided that the multiplexing capacity allows [18]. Moreover, as the amplicon size is less restrictive, generally shorter amplicons can be generated which poses an advantage, particularly when DNA of low quality (i.e., more degradation) is encountered [18]. Whether it be a length-based, or a sequencing-based genotyping method, there is a need for industry to play their part.

In summary: I envision that the research outcomes described in this thesis will motivate industry to put the opportunities that these highly mutable Y-STRs offer in practice of commercial kits.

Is it possible to achieve complete male differentiation and individual identification using Y-STRs?

The ultimate goal for the use of Y-STRs in forensics genetics would be to achieve complete male differentiation and thereby individual identification of males, being on par with their

autosomal counterparts. There are many forensic cases, particularly those of sexual assault, where due to the mixture of male perpetrator with female victim DNA in the evidence sample, the identification of the male perpetrator is technically impossible with standard autosomal STRs. To solve such cases, it would be the ultimate goal to derive individual identification from Y-chromosome analysis, thereby avoiding the problem of allele sharing existing with autosomal STR analysis in mixed samples. But can such goal ever be achieved? With the work described in this thesis and by almost doubling the number of RM Y-STRs compared to the time before this thesis work started, we made the step from approximately 25% to over 40% of father-son pair differentiation and more for other types of paternal relatives **(Chapter 4 and 6)**.

Another method that sequences over one hundred Y-STRs: approaches as CSYseq [16], could theoretically surpass the 50% father-son differentiation rate mark, albeit such performance of the method has yet to be demonstrated empirically and the high mutation rates estimated in the study need confirmation by independent analysis. There are many non-overlapping Y-STRs between CSYseq and RMplex; therefore, combining both methods might result in a further increase of father-son pairs differentiation, which yet needs to be practically demonstrated by empirical evidence. Moreover, there is a direct-to-consumers company that offers a test including 700 Y-STRs (https://www.familytreedna.com/products/y-dna). Although, it is unlikely that the latter would be compatible with forensic type of DNA samples.

However, the higher the differentiation rate achieved by a set of Y-STR markers becomes, the more difficult it would be to further enhance it. When only a small number of pairs is differentiated, the probability of a new mutation differentiating an additional pair is relatively high. However, when most pairs already have been differentiated it becomes more likely that new mutations occur in pairs that already had been differentiated by other Y-STRs than that they would differentiate a new pair. This phenomenon is visualized in Figure 1 and 2, here we used the simulated data that was described in **Chapter 6** for three genotyping assays: RMplex, Yfiler™ Plus PCR Amplification Kit, and PowerPlex® Y23 System. For each number of separating meioses (ranging from 1-50), we calculated the simulated differentiation rate and the mean number of separating mutations. Plotting these values against each other shows a clear pattern where for each incremental step in differentiation rate an increasingly larger step in mean mutation rate is needed. These figures demonstrate why it will be so challenging to ever reach 100% differentiation. The mean number of mutations between pairs is a value that approximates the sum of the mutation rates of all Y-STRs included in a kit. Therefore, to reach the mean value of one mutation between pairs, would require a kit including 100 Y-STRs with an average mutation rate of 10^{-2} mpg. As shown in Figure 1, such a kit would

differentiate only slightly over 60% of the father-son pairs. As following from Figure 1, differentiation 90% of father-son pairs would require an mean mutation rate per pair of ~2.3 (Figure 1), 95% could be reached with a mean value of ~3 (Figure 2), 99% with ~4.5 (Figure 2) and to approximate 100% would require a mean number of mutations per pair of ~8, i.e., 800 Y-STRs with an average mutation rate of 10^{-2} mpg.

It has to be noted that case-specific male individualization can already occur long before completely individualization of every man and in every lineage would be achieved. Mutations are highly stochastic and examples of male individualization could already be achieved with RMplex, as was shown in **Chapter 6.** The more the differentiation rate can be increased in the future, the more male individualization will become the rule, with undifferentiable male relatives becoming the exception. At the very least, it is evident that we have a long way to go before coming even close to complete male relative differentiation using the human Y-chromosome. Yet it may not be impossible. Other than Y-STRs mutations, in principle any nucleotide found on the Y-chromosome could become a single nucleotide variant (SNV) as the result of a mutation that will contribute to male relative differentiation. Although the mutation rates of Y-chromosomal SNVs (estimated to be in the order of 10^{-8} mpg [19]) are a lot lower compared to Y-STRs, the number of potential Y-SNV is orders of magnitude larger than that of Y-STRs. If we ever are going to approach complete male differentiation, and thus individual identification based on Y-chromosome evidence, to my view, it will be through the sequencing of as much of the human Y-chromosome as possible and combing all Y-STRs with all Y-SNVs included in the sequenced parts of the Y-chromosome, as could be possible by employing third generation sequencing technologies such as PacBio long range sequencing.

In summary: There is a long way to go before complete male differentiation solely based on Y chromosomes will become possible, perhaps 100% differentiation in each pair of males will prove simply impossible. Nevertheless, there still is a lot of room for improvement, in particular, when utilizing the capabilities of modern sequencing technology.

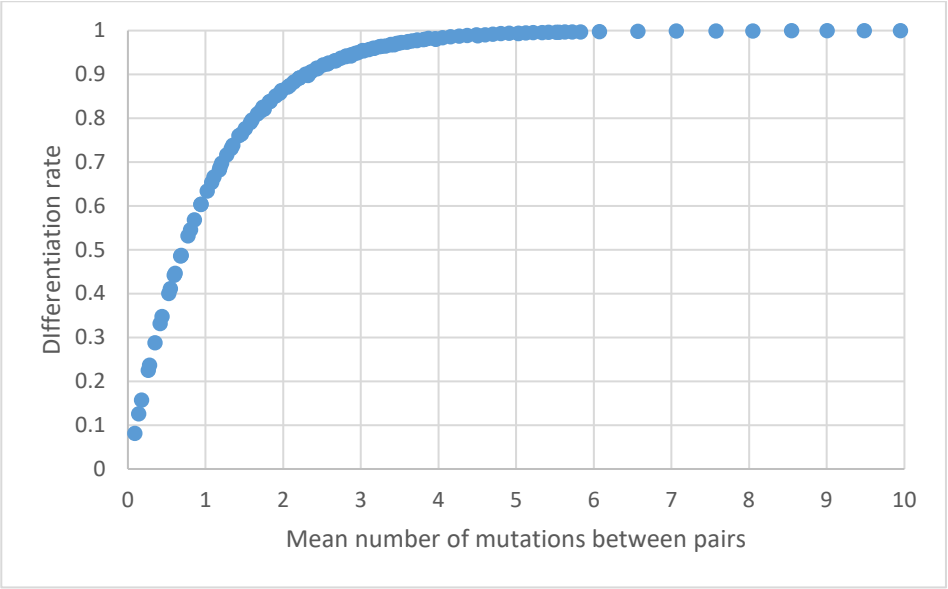


Figure 1: Correlation between the average number of mutations between pairs and the differentiate rate obtained with Y-STRs by simulating datasets of 100,000 pairs per generation in the range of 1-50 generations (with the full range of differentiation rates on the Y-axis).

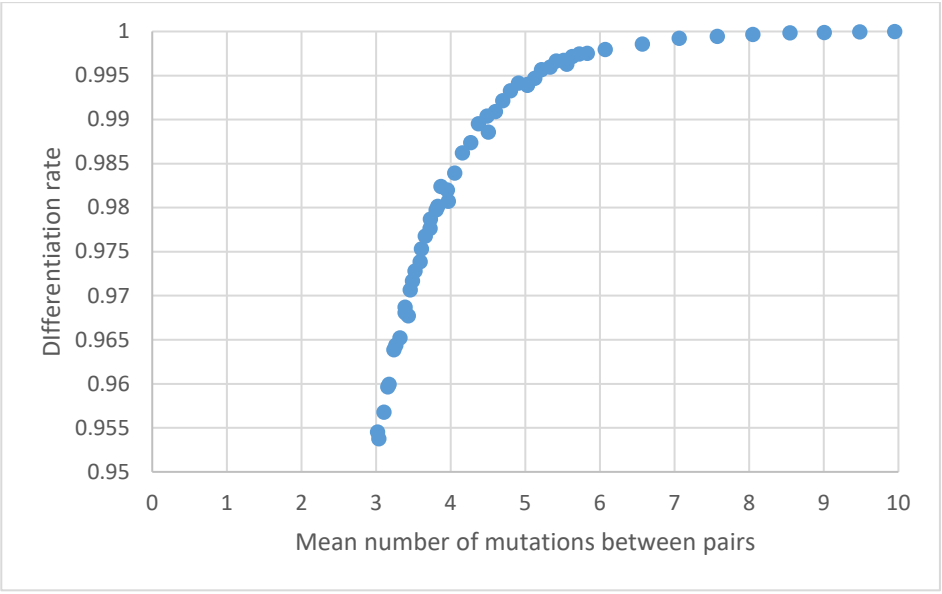


Figure 2: Correlation between the average number of mutations between pairs and the differentiate rate obtained by simulating datasets of 100,000 pairs per generation in the range of 1-50 generations (with the differentiation rate on the Y-axis ranging from 0.95 – 1).

Multiple male contributors

A relevant issue that had not yet been discussed in this thesis is the possibility of mixtures with DNA contributions from multiple male donors, for example as the result of sexual assault with multiple perpetrators, in addition to the DNA contribution of the female victim. Such mixtures would result in extra layers of complexity for both autosomal STR and Y-STR analysis. The severe consequences that taking such complex Y-STR profile interpretations too lightly was shown in a case report [20]. The report describes a case where a suspect was initially convicted for a multiple-attacker rape of two women. His conviction was based on a coincidental inclusion as contributor in a 17-locus Y-STR mixture. In the case report it was shown that based on only 17 standard Y-STR loci, all alleles from the wrongfully convicted suspect overlapped with an allele from one or the other of the two additional suspects. Further expanding the Y-STR set using a 23-locus system showed that the wrongfully convicted suspect carries two loci (including DYS576) that did *not* overlap with any alleles of the other two suspects; both alleles were also not present in the mixture, hence the suspect was exonerated [20]. Although in my opinion many things went wrong in the initial interpretation of the evidence that led to the conviction of this innocent man, the report does show the importance of using extended Y-STR marker sets and careful interpretation of the results.

RM Y-STRs may be especially useful in cases as described in the case report given their highly polymorphic nature [3], which reduces the probability of coincidental inclusions. Hence, the exclusion of suspected contributors to mixtures based on method like RMplex can be expected to be highly effective. Nevertheless, determining if a suspect (or one of his paternal relatives) contributed to a mixture, or if the matching alleles are rather the result of a coincidental inclusion, is more complex.

Here, deconvolution of the mixture and assigning all alleles to individual contributors would be of great importance. If DNA originating from two male contributors is present in a mixture at different ratios, it may be possible to assign alleles to the major and minor contributor, respectively, based on peak heights. This approach is already used for autosomal STRs in mixtures using probabilistic genotyping software [21]. Such software has yet to be developed for Y-STR profiles derived from mixtures with multiple male contributors. What could make this approach more difficult, especially when using a method like RMplex, is the presence of multi-copy loci and Y-STRs with high stutter peaks. Reliably assigning the correct alleles to the minor contributor may be complicated in such cases and some loci may require exclusion from the analyses because there is too much uncertainty in the genotyping result interpretation. Furthermore, when two or more

males have contributed equal amounts of DNA to the mixture, the peak heights cannot be used to deconvolute the mixture.

Another strategy that may be attempted is making use of the fact that all Y-STRs on the non-recombining portion of the Y-chromosome are linked and passed from generation to generation together. In consequence, certain alleles may be associated to one another. For example, if in Y-STR A allele 10 is found in 95% of the cases together with allele 20-22 in Y-STR B, while allele 14 is found in 95% of the cases together with allele 15-19 in Y-STR B, one could conclude that if in a mixture allele 10 and 14 are found for Y-STR A and allele 16 and 20 for Y-STR B, that it is most probable that allele 14 (A) and 16 (B) belong to one individual while allele 10 (A) and 20 (B) belong to the other contributor. This approach, which has some similarities with imputation as done in genome-wide association studies [22], may result in individual-specific haplotypes for at least the Y-STRs that show such distinct allelic associations. However, this approach would rely heavily upon the availability of large databases to infer the allele associations from.

Lastly, another, somewhat similar, approach would be to make use of knowledge about deeper evolutionary relationships (i.e., Y-SNP based haplogroups). It was shown previously, that using MPS and specialized software tools, high-resolution Y-haplogroup inference is possible from mixtures that contain multiple male contributors [23]. As males belonging to the same Y-haplogroup share a more recent common paternal ancestor than individuals belonging to different Y-haplogroups, it can be expected that there is less Y-STR haplotype diversity within Y-haplogroups than between Y-haplogroups [24]. Therefore, associations may exist between specific Y-haplogroups and allelic ranges of specific Y-STRs (**Chapter 5)**, which in turn could be used to link specific Y-STR alleles to individual contributors from mixtures to which multiple males contributed. However, also this approach would require large databases that on top of Y-STR data also containing Y-SNP based haplogroup data from the same individuals.

In summary: a lot more research will be needed before finding the best approach to handle mixed Y-STR profiles. Sufficiently large databases, combined with solid statistical frameworks may serve as good starting points to test the success of different approaches empirically.

Patrilineal investigative genetic genealogy

The potential of paternal investigative genetic genealogy was already discussed in **Chapter 6**, nevertheless I would like to emphasize on it more in this general discussion, as in my

opinion, this is where RM Y-STRs could have the largest societal impact. To reiterate: the approach that I envision would be to include the analysis of standard Y-STRs assays, mostly consisting of Y-STRs with moderate mutation in rates, in any forensic investigation on both the crime scene material and in new and historical (male reference sample of males included in the criminal offender DNA databases. The infrastructure of criminal offender DNA databases should easily be able to cope with these data as the nature is the same as that of autosomal STRs currently included in such forensic databases, boiling down to a string of (allelic repeat) numbers. Then for every case where no autosomal STR match was found between a crime scene sample and an individual stored in the criminal offender DNA database (because of a previous offense for which such male was convicted earlier), an additional query would be made to search for men in the criminal offender DNA database that have Y-STRs haplotypes with a maximum of one, or two variations compared to the crime scene sample (Y-match). Such Y-STR haplotypes would indicate a paternal familial relationship between the Y-match and the donor crime scene sample, while the Y-match could not have been the donor of the crime scene sample as shown by the absence of an autosomal STR match.

In some cases, there may be no Y-match present among the reference samples, which would mean the end of the line for such database-based investigation, at least in that point of time. However, given the conservative nature of Y-STR haplotypes, especially those consisting of Y-STRs with moderate mutation rates, it is expected that in many cases, one or multiple Y-matches would be present in the database. As soon as there is a Y-match, forensic investigators would have a starting point for further investigation. This is where the relevance of the work described in this thesis comes into play. Observing a Y-match indicates a high probability of sharing a common patrilineal ancestor with the donor of the crime scene sample. In the case of a recent common ancestor, this could lead to a breakthrough if combined with tactical police investigations, while if the Y-match and the donor of the crime scene sample share a common ancestor that lived hundreds of years in the past it will be unlikely to help in identifying the donor of the crime scene sample, unless perhaps when very thorough genealogical research would be conducted. RMplex was shown in **Chapter 6** to be remarkably suitable for exactly the purpose of making a distinction between closely related and distantly related male relatives.

Andersen and Balding developed models based on simulations to assess for different standard Y-STR genotyping kits the expected number of males with a fully matching Y-STR profile in the population [25]. For example, for PowerPlex® Y23 System their models show, depending on the exact parameter used in models, that a median of 13-25 males would share the same Y-STR haplotype in the population. If we consider that the Dutch criminal offender DNA database consist of ~350,000 reference profiles of which

~86% belong to males [26], there are approximately 300,000 male individuals included in the database. If each of those males in the database would have 13 other males sharing their standard Y-STR haplotype, it would mean that using Y-STRs, the number of males indirectly included (i.e., via their relatives directly included) in the database would reach 4 million (almost 50% of the male Dutch population).

This estimation assumes that all 300,000 men already in the Dutch criminal offender DNA database are unrelated to each other, which likely is not the case based on close or distant paternal relationship. Furthermore, it would be assumed that the 300,000 males in the database form an unbiased representation of the Dutch male population, which probably also is not the case. Then again, this estimation assumed the lowest possible number of a median of 13 matching Y-STR haplotypes per male, according to Andersen and Balding [25], which may be an underestimation. Moreover, this estimation only considers full Y-STR matches, while a near match with one or two differences compared to the crime scene sample may occur, even in close relatives. Considering all these factors together, it may not be so unrealistic that about half of the male Dutch population would have some degree of patrilineal representation in the current Dutch criminal offender DNA database. Such information could be harvested to find unknown male offenders via their paternal relatives in the database if Y-STR profiles of the males in the database would be available.

Assuming the cost to generate standard Y-STR profiles to be approximately €50 per sample, it would be possible to fully upgrade the Dutch DNA database with Y-STR profiles for €15.000.000. By no means cheap, but relative to the total spending of the Dutch ministry of Justice and Security, which was almost 15 billion euros in 2021, a rather negligible amount. The benefit would be that many unsolved criminal cases, including those of high severity such as rape and murder cases, could find a new opening for further investigation. Every case that can be solved could give some form of closure to the victims, or their surviving relatives, and prevents further crime at least for the number of years the convicted offender is imprisoned and thus unable to continue his criminal behavior. Moreover, once included in the criminal offender database with his autosomal DNA profile, identifying that individual a second time could be simply done using autosomal STR typing.

To my view, 15 million euros, or even 100 million euros would be a relatively small price to pay to avoid the suffering of innocent victims of crimes and to bring offenders of serious crimes to justice. It is my sincere hope that this thesis will contribute to making those in political power realize the potential opportunity of the use of Y-STRs when implemented on the level of national criminal offender DNA databases that until

now has been cast aside, despite the Netherlands Forensic Institute having made similar arguments before.

In summary: Many more crimes could be solved or prevented if the patrilineal genetic genealogy approach were to be adopted. What is still needed is political willpower and resources. Perhaps most importantly there is a need for a thorough consideration of the potential negative ethical and societal effects that such an approach could have. Although, such considerations are beyond the scope of this thesis, they are crucial, as is a solid legal framework to decide when it is appropriate and proportional to resort to such means of investigation.

Investigative genetic genealogy using SNPs vs. using Y-STRs

One could argue that it may be more effective to move towards investigative genetic genealogy using autosomal SNPs (SNP-IGG), as this method has proven to be a very powerful in solving cold cases [27]. And there are definitely certain advantages of using this approach over using Y-STRs (Y-IGG):

1) Obviously, Y-STRs are completely useless when the donor of a crime scene sample is female, whereas for SNP-IGG the sex of the donor is irrelevant.
2) SNP-IGG can point to single individuals, or at most to a group of full siblings, whereas Y-IGG, depending on the case, may leave a group of potential contributors to a crime scene trace.
3) SNP-IGG can use public databases which are already existing, databases for Y-IGG would mostly have to be build up from the ground.
4) SNP-IGG can potentially also be used for individual identification, Y-IGG ideally required autosomal STR typing to unequivocally confirm the identity of the donor to the crime scene trace.

However, Y-IGG also advantages over SNP-IGG:

1) SNP-IGG is not possible in female-male mixtures as often encountered in sexual assault cases, which is a type of high-frequency crime worldwide. Y-STRs are particularly suitable for these type of cases.
2) SNP-IGG relies on databases where the public can upload their data and that are typically owned by commercial parties. For forensic investigators it is not ideal to rely on a database that is controlled and populated by third-parties as they could be tempered with. Y-STRs, however, are highly compatible with current criminal offender databases and could be employed in the same way

by the same agencies outside of the reach of the public and industrial parties.

3) Forensic laboratories are very experienced in working with STRs, SNP-IGG required either SNP arrays, or high throughput sequencing methodologies which require a specialized skillset. Moreover, these technologies are likely less suitable for forensic type of samples (i.e., low template, mixtures, etc.) compared to STR typing.
4) Y-IGG only needs a single match to form a starting point for further investigation, SNP-IGG typically relies on at least two matches, one from the paternal and one from the maternal side of the donor of the crime scene stain.
5) The databases used for SNP-IGG are biased, they work well in the US as there are many profiles of people living in the US included, this may be less the case e.g., in European countries, more so in less developed counties such as in Africa.

In summary: It is my opinion that fighting crime and solving cases is the responsibility of governmental agencies with a mandate to do so. If SNP-IGG were to be applied, it is my believe that it should be those agencies that perform the analysis and maintain the database for this purpose. Public databases populated by individuals interested in their family history or other aspects of their genetic heritage should serve the interest of their customers and not those of the authorities. Ideally, I would envision Y-IGG and SNP-IGG existing side-by-side; however, I see Y-IGG more as the low-hanging fruit that could have a large impact with a relatively small change to the status quo.

Conclusions and outlook

Chapter 2 showed that it is possible to scan Y-chromosome sequences and to predict which Y-STRs are likely to display above average mutation rates, and follow up by empirical confirmation. Future studies may employ a similar approach based on a variety of Y-chromosome references sequences from as diverse populations as possible to find more Y-STRs with increased mutation rates to overcome the limitations seen with the markers identified with (and before) this thesis work. The number of repetitive units is the main driver of mutation rates. Therefore, making a candidate selection based on multiple Y chromosomal sequences and by using, for example, median values of the repeat numbers, stochastic variations that can result from predictions based on a single Y chromosome can be mitigated. Moreover, as STRs with high mutation rates are typically the most polymorphic, candidate selection approach that was described here could also be used to identify targets for expanded, highly discriminatory autosomal STR kits, which may be useful to, for example, deconvolute DNA mixtures. **Chapter 2** also contains preliminary evidence that the sequence of the repeated motif has an influence on the mutation rate. In particular the AAAG-motif and tetranucleotide repeats and the AAG-motif in trinucleotide repeats appear to be dominant in Y-STRs that express high mutability. Future studies, e.g., using yeast models, may shed more light on the evolutionary mechanisms that drive these motif sequence specific differences.

Chapter 3 described an approach to iteratively develop fragment length based multiplex PCR assays, the approach could be applied to other STR multiplex assays targeting autosomal DNA, or completely different organisms. The developmental validation showed that, albeit not at the same level as industry developed assays, RMplex is able to deliver good results from the most encountered types of forensic material. As a result of the validation, RMplex can now be applied on case work within the Netherlands Forensic Institute. Having developed an efficient genotyping method also invites other researchers to putting the new RM Y-STRs to the test and to see if similarly high mutation rates can be estimated using their local population of close male relatives. There is a need for a lot more reference data, especially for the newly proposed RM Y-STRs from **Chapter 2**, RMplex provides a method to produce such data efficiently. Future methods may resort to MPS rather than CE, as the latter has mas many restrictions in the number of Y-STRs that can be combined in a single assay. That

Chapter 4 demonstrated the performance of RMplex for the first time in a completely independent set of close male relatives (i.e., father-son pairs and brothers). Here, for the first time, it was shown that RMplex has the capability to differentiate over 40% of the fathers from their sons. Moreover, **Chapter 4** provides novel consensus

mutation rates estimated for all Y-STRs included in RMplex and in Yfiler™ Plus PCR Amplification Kit by combining the newly generated data with data from previously published studies focusing on Y-STR mutation rate using father-son pairs. The overall consensus estimates are based on large numbers of father-son pairs, mostly from diverse populations and therefore provide the best estimations of mean locus-specific mutation rates up till date. However, as discussed previously in this chapter, mutation rate estimations should not be regarded as an objective truth that hold true in every scenario. There is a need for more studies and to keep refining the mutation rate estimates as more reference data becomes available.

Chapter 5 showed the first application of RMplex on a sample set with close relatives of non-European background. A noteworthy observation was the significantly higher differentiation rate of Japanese father-son pairs compared to those from Austria as described in **Chapter 4.** This difference was mostly driven by a total of six Y-STRs that showed significantly higher mutation rates in the Japanese men compared to the multi-study consensus mutation rates as described in **Chapter 4**. Acknowledging the relatively small number of Japanese father-son pairs that were included in the study, some reservation is required before drawing definitive conclusion, at least until the results will be replicated by larger scale studies. Nevertheless, **Chapter 5** also delivered preliminary evidence that the different mutation rate estimated between populations may, in part, be the result of different allelic distributions among different haplogroups. If the existence of such effects can be further supported by independent future research on different populations / haplogroups it could have major implications on how we look at mutation rate estimation. The following different levels of mutation rate estimation could be considered:

- Locus-specific mutation rates are the values that are now typically reported and the represent the average mutation rates amongst different studies.
- Population-specific mutation rates would represent the average mutation rate estimates observed in a clearly defined (e.g., geographically, or culturally) populations.
- Haplogroup-specific mutation rates would represent the average mutation rate estimates within specific haplogroups. An important requirement would be a consensus on the level of haplogroup resolution at which such mutation rates should be best determined.
- Size-based allele-specific mutation rates would take into account the fact that longer alleles are more likely to mutate than shorter alleles and would, therefore, provide a separate mutation rate estimate for each observed allele.

- Sequence-based allele-specific mutation rates would, in my opinion, be the most accurate way to look at mutation rate. Each individual repeat stretch within each locus would obtain its own empirically derived mutation rate. This approach would disregard biases that exist in all other previously suggested levels of mutation rate estimation.

Using any of these types of mutation rate estimations other than the locus-specific mutation rate would likely result in more accurate and individualized estimations; however, it would also require generating an order of magnitude more empirically data than is available currently. And it would require new guidelines on how to interpret and report observed mutations. Moving away from locus-specific mutation rates would also suggest a need to move away from genotyping-kit specific differentiation rates. In such a future scenario it could become possible to estimate haplotype-specific differentiation rates.

Chapter 6 described the application of RMplex in a large number of male pedigrees with different characteristics in regard of depth, number of included males and biogeographic ancestry. Here, it was demonstrated for the first time, how effectively RMplex can differentiate closely and more distantly related males and that complete male individualization, in cases, was possible. Moreover, it was shown the Y-STRs with high mutation rates are the most effective type of markers to predict the level of patrilineal consanguinity. It can be concluded that to obtain an impression of the mutability of a yet uncharacterized Y-STRs, using pedigrees can provide a more cost-efficient alternative to the typing of many father-son pairs. Although estimating mutation rates through pedigrees comes with some caveats, the overall trends between the two approaches were consistent. The capacity that RMplex provides to reduce the number of shared haplotypes between individuals within a pedigree of closely related males opens up new avenues for forensic genetics, like the use of patrilineal investigative genetic genealogy as extensively discussed in **Chapter 6** and earlier in this chapter. This new application of Y-STRs to forensic genetics is expected to increase in relevance and performance when more Y-STRs with increased mutation rates will become available and will be combined with already identified Y-STRs in more extensive MPS-based Y-STR genotyping assays.

Closing remarks

Taken all together, the work described in this thesis represents a major step, but nonetheless just one step towards maximization of the applicability of Y-STRs to improve

forensic genetics and beyond. I have no doubt that there is much more to gain in the future, which will require efforts from scientists, forensic practitioners, industry, politicians, ethicists, genealogist and other stakeholders. I hope to see the day that the use of human Y chromosome polymorphisms such as Y-STRs but also Y-SNPs in routine forensic genetic casework develops from a small niche application to a fundamental cornerstone of the genetic toolbox used to fight crime and injustice.

References

1. Kayser, M., *Forensic use of Y-chromosome DNA: a general overview.* Human Genetics, 2017. **136**(5): p. 621-635.
2. Adnan, A., et al., *Improving empirical evidence on differentiating closely related men with RM Y-STRs: a comprehensive pedigree study from Pakistan.* Forensic Science International: Genetics, 2016. **25**: p. 45-51.
3. Ballantyne, K.N., et al., *Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications.* The American Journal of Human Genetics, 2010. **87**(3): p. 341-353.
4. Brinkmann, B., et al., *Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat.* The American Journal of Human Genetics, 1998. **62**(6): p. 1408-1415.
5. Eckert, K.A. and S.E. Hile, *Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome.* Molecular Carcinogenesis, 2009. **48**(4): p. 379-388.
6. Ellegren, H., *Microsatellites: simple sequences with complex evolution.* Nature Reviews Genetics, 2004. **5**(6): p. 435.
7. Kayser, M., 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.* The American Journal of Human Genetics, 2000. **66**(5): p. 1580-1588.
8. Kelkar, Y.D., et al., *The genome-wide determinants of human and chimpanzee microsatellite evolution.* Genome Research, 2008. **18**(1): p. 30-38.
9. Willems, T., et al., *Population-scale sequencing data enable precise estimates of Y-STR mutation rates.* The American Journal of Human Genetics, 2016. **98**(5): p. 919-933.
10. Xu, X., et al., *The direction of microsatellite mutations is dependent upon allele length.* Nature Genetics, 2000. **24**(4): p. 396.
11. Ballantyne, K.N., et al., *A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages.* Forensic Science International: Genetics, 2012. **6**(2): p. 208-218.
12. Kayser, M., et al., *A comprehensive survey of human Y-chromosomal microsatellites.* The American Journal of Human Genetics, 2004. **74**(6): p. 1183-1197.
13. Benson, G., *Tandem repeats finder: a program to analyze DNA sequences.* Nucleic Acids Research, 1999. **27**(2): p. 573-580.
14. Nurk, S., et al., *The complete sequence of a human genome.* Science, 2022. **376**(6588): p. 44-53.
15. D'Angelo, O., et al., *Assessing non-LUS stutter in DNA sequence data.* Forensic Science International: Genetics, 2022. **59**: p. 102706.

16. Claerhout, S., et al., *CSYseq: The first Y-chromosome sequencing tool typing a large number of Y-SNPs and Y-STRs to unravel worldwide human population genetics.* PLOS Genetics, 2021. **17**(9): p. e1009758.
17. Claerhout, S., et al., *A game of hide and seq: Identification of parallel Y-STR evolution in deep-rooting pedigrees.* European Journal of Human Genetics, 2019. **27**(4): p. 637.
18. Børsting, C. and N. Morling, *Next generation sequencing and its applications in forensic genetics.* Forensic Science International: Genetics, 2015. **18**: p. 78-89.
19. Xue, Y., et al., *Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree.* Current Biology, 2009. **19**(17): p. 1453-1457.
20. Hampikian, G., et al., *Case report: Coincidental inclusion in a 17-locus Y-STR mixture, wrongful conviction and exoneration.* Forensic Science International: Genetics, 2017. **31**: p. 1-4.
21. Bleka, Ø., G. Storvik, and P. Gill, *EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts.* Forensic Science International: Genetics, 2016. **21**: p. 35-44.
22. Marchini, J. and B. Howie, *Genotype imputation for genome-wide association studies.* Nature Reviews Genetics, 2010. **11**(7): p. 499-511.
23. Ralf, A., et al., *Forensic Y-SNP analysis beyond SNaPshot: High-resolution Y-chromosomal haplogrouping from low quality and quantity DNA using Ion AmpliSeq and targeted massively parallel sequencing.* Forensic Science International: Genetics, 2019. **41**: p. 93-106.
24. Larmuseau, M.H.D., et al., *Recent radiation within Y-chromosomal haplogroup R-M269 resulted in high Y-STR haplotype resemblance.* Annals of human genetics, 2014. **78**(2): p. 92-103.
25. Andersen, M.M. and D.J. Balding, *How convincing is a matching Y-chromosome profile?* PLOS Genetics, 2017. **13**(11): p. e1007028.
26. Meulenbroek, L., *DNA zoekmachine.* 2021.
27. Dowdeswell, T.L., *Forensic genetic genealogy: A profile of cases solved.* Forensic Science International: Genetics, 2022. **58**: p. 102679.

Addendum

Summary

Short tandem repeats located at the non-recombining portion of the human Y chromosome (Y-STRs) have been used in the fields of forensic genetics, genealogy and anthropology for decades. Their haploid and slowly mutating nature made them especially suitable genetic markers to investigate male human evolution and migration; the same features made them also very suitable to identify unknown patrilineal relationships. However, as evidence in court the applicability was limited, as many males sharing a common patrilineal ancestor would show the exact same haplotypes. About a decade ago this started to shift with the identification of the first set of rapidly mutating (RM) Y-STRs, which are Y-STRs that mutate at least once every 100 generations, which an increase of about tenfold compared to the average Y-STR mutation rate.

In **Chapter 1** the usage of Y-STRs in forensic genetics is placed in the larger context starting from the earliest applications of forensic sciences, through the discovery of DNA and major technological advances in DNA technology, to the current state-of-the-art type of forensic DNA analysis. This chapter further introduces the beginning of interest in Y-STRs from the field of forensic genetics and how the usage of Y-STRs has evolved over time. The societal relevance is exemplified by describing a high-profile case in the Netherlands that could eventually be solved by using the unique characteristics of Y-STRs. Lastly, **Chapter 1**, formulates the different aims of this PhD thesis.

Chapter 2 describes a study in which we sought to identify more RM Y-STRs than were available at the time. To achieve this goal we developed a system to assign a 'mutability prediction score' to any Y-STR that could be found on the human Y-chromosome. The system was based on molecular characteristics that set previously identified RM Y-STRs apart from Y-STRs with lower mutation rates. Using this approach 27 candidate RM Y-STRs with high mutability prediction scores were ascertained. Importantly, these 27 candidates RM Y-STRs had not been included in previous large-scale mutation rate studies. By genotyping a total of over 1,600 father and their sons we could confirm that, indeed 12 out of these 27 candidates, showed a mutation rate that would qualify them as RM Y-STRs. Many other candidates showed mutations rates that were above average, but not sufficiently high to classify them as RM Y-STRs. Based on the 647 mutations that had been observed we performed analyses to confirm the role of previously suggested factors that could influence (Y-)STR mutability. Lastly, we found support for the hypothesis that the repeat motif sequence plays an important part in the mutability of Y-STRs.

Chapter 3 deals with the development and validation of a new genotyping method, termed RMplex, to analyze all currently known RM Y-STRs and four additional Y-

STRs with increased mutation rate estimates. A novel approached was used to optimize the distribution of the different Y-STRs over two multiplexes and four available fluorescent dye channels. The multiplex was, in principle, designed to be applied with CE. However, in this study we used the same multiplex conditions to sequence the STRs using an Illumina Miseq sequencer using a small number of individuals. These sequencing data were used to propose a repeat number nomenclature for the novel markers, which can be used as an convention to homogenize the data that will be produced from future studies targeting these Y-STRs. The validation demonstrated that the results obtained with RMplex are highly repeatable and reproducible. Moreover, full profiles could be obtained with as little as 100 pg of genomic input DNA. When using 1 ng of male input DNA, full profiles could be obtained in male:female DNA mixtures with a ratio of 1:50. When DNA was degraded below a mean fragment size of 400 bp, in particular, the larger Y-STRs started to show drop-out. RMplex was shown in this chapter to be suitable to be applied to the most common types of forensic samples; however, the technical performance could not compare to some of the latest generation commercial Y-STR genotyping kits.

Chapter 4 describes the first application of RMplex on a large number of 499 father-son pairs and 92 brothers. The performance of RMplex was also put into context by comparing the results to those of Yfiler™ Plus PCR Amplification Kit (Yfiler Plus), which is the state-of-the-art commercial Y-STR genotyping kit. Using RMplex a total of 289 mutations were observed the 499 father-son pairs, while Yfiler Plus detected 76 among 530 father-son pairs. Based on these results we estimated locus-specific mutation rates, no significant differences were observed after correcting for multiple testing. Moreover, we provided consensus mutation rates estimates by combining the newly described estimates with father-son pair based mutation rate estimates derived from literature for all 49 Y-STRs included in either of both applied genotyping methods. **Chapter 4**, described the first empirical based differentiation rate estimate for RMplex which was 42% for RMplex, while Yfiler Plus only differentiated 13% of the father-son pairs. Brothers were differentiated in 33% an 62% of the cases with Yfiler Plus and RMplex, respectively. Both methods managed to differentiate all unrelated males (i.e., the fathers) from each other, with the exception of one pair that had shared haplotypes for both methods. This is likely explained by both males being patrilineally related rather than being unrelated, due to anonymization we could not verify this assumption.

Chapter 5 described the first study where RMplex is applied on father-son pairs from outside of Europe. This study genotyped 178 Japanese father-son pairs with both RMplex and Yfiler Plus and found a total of 138 mutations using RMplex and 29 with Yfiler Plus. A remarkable result was that a total five Y-STRs showed significantly higher mutations rates in the Japanese father-son pairs compared to the reference mutation rate

estimates. As a result RMplex delivered a significantly higher (p-value 0.0179) differentiation rate of 52% compared to the study described in **Chapter 4** where 42% of the father-son pairs were differentiated. In contrast, Yfiler Plus differentiated 13% of the father-son pairs which is in agreement with the study described in **Chapter 4**. To evaluate if the increased mutation rate estimates could be explained by evolutionary differences specific to the Japanese population, we developed an assay to detect the most common Y-haplogroups observed in the Japanese population. Consequently, we compared the haplogroup-specific allele frequencies of one of the Y-STRs with a significantly higher mutation rate (DYS712). The results indicated that indeed the Japanese individuals on average had longer alleles for this Y-STR; moreover, one haplogroup showed particularly long alleles. Although the sample size in this study was limited, it shows how based on haplogroup composition can different populations can display different locus-specific mutation rate estimates.

In **Chapter 6**, we moved the focus beyond father-son pairs, by typing pedigrees belonging to three different cohorts using RMplex. These cohorts differed from one another in several aspects, one was mostly characterized by close relatives of South Asian origin, while another one contained mostly (very) distantly related males from Western Europe and the last cohort covered a wide range of relationship including closely and more distantly related male relatives from a rural village in the Netherlands. Additionally, the last described cohort was also typed using Yfiler Plus. The results from the study show the high capability of RMplex for male differentiation, father-son paired were differentiated in 43% of the cases using RMplex, in line with the results from **Chapter 4**. Males separated by four generations (e.g., first cousins) could be differentiated in 84% of the cases, while and over 95% of the males separated by six generations could be differentiated by RMplex. Additionally, the pedigrees were used to estimate the mutation rates, an approach of which both the strengths and the weaknesses were discussed in Chapter 6. Moreover, it was shown that RMplex can be used to predict the degree of consanguinity a lot more precise than when using standard Y-STR genotyping kits like Yfiler Plus. Lastly, this chapter discusses how these finding could be practically applied in forensic genetics to identify previously unknown persons of which the DNA was found at crime scenes.

The general discussion in **Chapter 7** provides a combined discussion of the results and conclusions from all previous chapters. Moreover, it provides several recommendation on how to further improve and expand the applicability of Y-STRs in the field of forensic genetics.

Samenvatting

Short tandem repeats op het niet-recombinerende deel van het menselijke Y-chromosoom (Y-STR's) worden al tientallen jaren gebruikt in de forensische genetica, genealogie en antropologie. Hun haploïde en langzaam muterende aard maakte hen bijzonder geschikte genetische markers om menselijke evolutie en migratie vanuit het oogpunt van de man te onderzoeken. Dezelfde kenmerken maken ze ook geschikt om onbekende patrilineaire relaties te identificeren. Echter, als bewijs in de rechtbank was de toepasbaarheid beperkt, aangezien veel mannen die een gemeenschappelijke patrilineaire voorouder delen exact dezelfde haplotypes zouden vertonen. Ongeveer tien jaar geleden begon dit te verschuiven met de identificatie van de eerste set snel muterende (RM) Y-STR's, deze groep Y-STR's muteert minstens eens per 100 generaties, dat is ongeveer tien keer zo vaak als de gemiddelde Y-STR.

In **Hoofdstuk 1** wordt het gebruik van Y-STR's in de forensische genetica in een bredere context geplaatst, beginnend bij de vroegste toepassingen van forensische wetenschappen, via de ontdekking van DNA en belangrijke technologische vooruitgang in DNA-technologie, tot de huidige state-of-the-art vorm forensische DNA-analyse. Dit hoofdstuk introduceert verder het begin van de belangstelling voor Y-STR's in het veld van forensische genetica en hoe het gebruik van Y-STR's in de loop van de tijd is geëvolueerd. De maatschappelijke relevantie wordt geïllustreerd door een spraakmakende zaak in Nederland te beschrijven die uiteindelijk werd opgelost door gebruik te maken van de unieke kenmerken van Y-STR's. **Hoofdstuk 1** formuleert tenslotte de verschillende doelstellingen van dit proefschrift.

Hoofdstuk 2 beschrijft een studie waarin we probeerden meer RM Y-STR's te identificeren dan er op dat moment beschikbaar waren. Om dit doel te bereiken hebben we een systeem ontwikkeld om een 'mutatie predictie score' toe te kennen aan elke Y-STR die gevonden kan worden op het menselijke Y-chromosoom. Het systeem is gebaseerd op moleculaire kenmerken die eerder geïdentificeerde RM Y-STR's onderscheiden van Y-STR's met lagere mutatiesnelheden. Met behulp van deze benadering werden 27 kandidaat-RM Y-STR's met hoge mutatiesnelheid voorspellingsscores verkregen. Belangrijk is dat deze 27 kandidaten RM Y-STR's niet waren opgenomen in eerdere grootschalige mutatiesnelheidsstudies. Door in totaal meer dan 1.600 vaders en hun zonen te genotyperen, konden we bevestigen dat inderdaad 12 van deze 27 kandidaten een mutatiesnelheid vertoonden die hen zou kwalificeren als RM Y-STR's. Veel andere kandidaten vertoonden mutatiesnelheden die bovengemiddeld hoog waren, maar niet hoog genoeg om ze als RM Y-STR's te classificeren. Op basis van de 647 mutaties die zijn waargenomen voerden we analyses uit om de rol te bevestigen van eerder gesuggereerde

factoren die (Y-) STR-mutabiliteit zouden kunnen beïnvloeden. Ten slotte vonden we ondersteuning voor de hypothese dat de sequentie van het repeterende motief een belangrijke rol speelt in de mutabiliteit van Y-STR's.

Hoofdstuk 3 behandelt de ontwikkeling en validatie van een nieuwe genotyperingsmethode, RMplex genaamd. Met behulp van deze methode kunnen alle op dit moment bekende RM Y-STR's en vier aanvullende Y-STR's met verhoogde mutatiesnelheden worden geanalyseerd. In de studie is een nieuwe aanpak toegepast om de verdeling van de verschillende Y-STR's over twee multiplexen en vier beschikbare kleurkanalen te optimaliseren. De multiplex is in principe ontworpen om in combinatie met capillaire elektroforese (CE) te worden toegepast. In deze studie hebben we echter dezelfde multiplexcondities gebruikt om de STR's te sequencen met behulp van een Illumina Miseq-sequencer in een klein aantal individuen. Deze sequentiegegevens werden gebruikt om een herhalingsnummer nomenclatuur voor de nieuwe markers voor te stellen, deze kan worden gebruikt als een conventie om de gegevens die zullen worden geproduceerd in toekomstige studies die deze Y-STR's gebruiken te homogeniseren. De validatie toonde aan dat de met RMplex verkregen resultaten zeer herhaalbaar en reproduceerbaar zijn. Bovendien konden volledige profielen worden verkregen met slechts 100 pg genomisch input-DNA. Bij gebruik van 1 ng mannelijk input-DNA konden volledige profielen worden verkregen in man-vrouw DNA-mengsels met een verhouding van 1:50. In gedegradeerde DNA monsters met een gemiddelde fragmentgrootte van minder dan 400 bp, begonnen de langere Y-STR fragmenten uitval te vertonen. In dit hoofdstuk is aangetoond dat RMplex geschikt is om te worden toegepast op de meest voorkomende soorten forensische monsters; de technische prestaties waren echter niet te vergelijken met sommige van de nieuwste generatie commerciële Y-STR-genotyperingskits.

Hoofdstuk 4 beschrijft de eerste toepassing van RMplex in een groot aantal van 499 vader-zoon paren en 92 broers. De effectiviteit van RMplex werd ook in context geplaatst door de resultaten te vergelijken met die van Yfiler™ Plus PCR-amplificatiekit (Yfiler Plus), de meest geavanceerde commerciële Y-STR-genotyperingskit. Met behulp van RMplex werden in totaal 289 mutaties waargenomen in de 499 vader-zoon-paren, terwijl Yfiler Plus er 76 detecteerde in 530 vader-zoon-paren. Op basis van deze resultaten maakte we een inschatting van locus-specifieke mutatiesnelheden, er werden geen significante verschillen waargenomen na correctie voor meervoudige testen. Daarnaast hebben we schattingen van de consensusmutatie snelheden geleverd door de nieuw beschreven resultaten te combineren met op vader-zoonpaar gebaseerde schattingen van de mutatiesnelheid vanuit de literatuur voor alle 49 Y-STR's die zijn opgenomen in een van beide toegepaste genotyperingsmethoden. **Hoofdstuk 4** beschrijft ook de eerste

empirisch gebaseerde schatting van het differentiatiepercentage voor RMplex, dat 42% was voor RMplex, terwijl Yfiler Plus slechts 13% van de vader-zoonparen differentieerde. Broers werden onderscheiden in 33% en 62% van de gevallen met respectievelijk Yfiler Plus en RMplex. Beide methoden slaagden erin om alle niet-verwante mannen (d.w.z. de vaders) van elkaar te onderscheiden, met uitzondering van één paar dat met beide methoden overeenkomstige haplotypes vertoonden. Dit kan waarschijnlijk worden verklaard doordat beide mannen patrilineair verwant zijn; echter vanwege anonimisering konden we deze veronderstelling niet verifiëren.

Hoofdstuk 5 beschrijft de eerste studie waarin RMplex wordt toegepast op vader-zoon paren van buiten Europa. In deze studie werden 178 Japanse vader-zoonparen gegenotypeerd met zowel RMplex als Yfiler Plus en werden in totaal 138 mutaties gevonden met RMplex en 29 met Yfiler Plus. Een opmerkelijk resultaat was dat in totaal vijf Y-STR's significant hogere mutatie snelheden vertoonden in de Japanse vader - zoonparen vergeleken met de schattingen van de referentie mutatiesnelheden. Als resultaat leverde RMplex een significant hoger (p-waarde 0,0179) differentiatiepercentage van 52% vergeleken met de studie beschreven in **Hoofdstuk 4**, waar 42% van de vader-zoon paren werden gedifferentieerd. Yfiler Plus differentieerde daarentegen 13% van de vader-zoonparen, wat in overeenstemming is met de studie beschreven in **Hoofdstuk 4**. Om te evalueren of de verhoogde schattingen van de mutatiesnelheid verklaard kunnen worden door evolutionaire verschillen die specifiek zijn voor de Japanse populatie, hebben we een test ontwikkeld voor het detecteren van de meest voorkomende Y-haplogroepen in de Japanse bevolking. Vervolgens vergeleken we de haplogroep-specifieke allelfrequenties van een van de Y-STR's met een significant hogere mutatiesnelheid (DYS712). De resultaten gaven aan dat inderdaad de Japanse individuen gemiddeld langere allelen hadden voor deze Y-STR; bovendien vertoonde één haplogroep opmerkelijk lange allelen. Hoewel de grote van de steekproef in dit onderzoek beperkt was, laat het zien hoe verschillende populaties op basis van de samenstelling van de haplogroep, verschillende locus-specifieke schattingen van de mutatiesnelheid kunnen vertonen.

In **Hoofdstuk 6** hebben we de focus verder verlegd van vader-zoon paren naar stambomen. Deze stambomen behoorde tot drie verschillende cohorten en zijn met behulp van RMplex getypeerd. Deze cohorten verschilden in verschillende opzichten van elkaar, de ene werd voornamelijk gekenmerkt door naaste verwanten van Zuid-Aziatische afkomst, terwijl een andere voornamelijk (zeer) verre verwante mannen uit West-Europa bevatte en de laatste cohort besloeg een breed scala aan relaties, waaronder nauw en meer verre verwante mannelijke familieleden uit een plattelandsdorp in Nederland. Bovendien werd het laatst beschreven cohort ook getypt met Yfiler Plus. De resultaten van

de studie laten zien dat het hoge vermogen van RMplex voor mannelijke differentiatie, vader-zoon paren werden gedifferentieerd in 43% van de gevallen met RMplex, in overeenstemming met de resultaten uit **Hoofdstuk 4**. Mannen gescheiden door vier generaties (bijv. neven) konden in 84% van de gevallen worden onderscheiden, terwijl en meer dan 95% van de mannen gescheiden door zes generaties door RMplex konden worden onderscheiden. Daarnaast werden de stambomen gebruikt om de mutatiesnelheden in te schatten, een benadering waarvan zowel de sterke als de zwakke punten werden besproken in H**oofdstuk 6**. Bovendien werd aangetoond dat RMplex kan worden gebruikt om de mate van bloedverwantschap veel nauwkeuriger te voorspellen dan wanneer met behulp van standaard Y-STR genotyperingskits zoals Yfiler Plus. Ten slotte wordt in dit hoofdstuk besproken hoe deze bevindingen praktisch kunnen worden toegepast in de forensische genetica om voorheen onbekende personen te identificeren waarvan het DNA op plaats delict is gevonden.

De algemene discussie in **Hoofdstuk 7** geeft een gecombineerde bespreking van de resultaten en conclusies uit alle voorgaande hoofdstukken. Bovendien geeft het verschillende aanbevelingen over hoe de toepasbaarheid van Y-STR's op het gebied van forensische genetica verder kan worden verbeterd en uitgebreid.

List of publications

1. **Ralf. A.**, Montiel González, D., Zandstra, D., van Wersch, B., Kousouri, N., de Knijff, P., … & Kayser, M. (2022). Large-scale pedigree analysis highlights rapidly mutating Y-chromosomal short tandem repeats for differentiating patrilineal relatives and predicting their degrees of consanguinity. *Human Genetics.* **(This thesis)**

2. Otagiti, T., Noriko, S., Asamura, H., Parvanova, E., Kayser, M., & **Ralf, A.** (2022). RMplex reveals population differences in RM Y-STR mutation rates and provides improved father-son differentiation in Japanese. *Forensic Science International: Genetics. 61,* 102766. **(This thesis)**

3. Xavier, C., de la Puente, M., Mosquera-Miguel, A., Freire-Aradas, A., Kalamara, V., **Ralf, A.**, ... & VISAGE Consortium. (2022). Development and inter-laboratory evaluation of the VISAGE Enhanced Tool for Appearance and Ancestry inference from DNA. *Forensic Science International: Genetics 61,* 102779.

4. Neuhuber, F., Dunkelmann, B., Grießner, I., Helm, K., Kayser, M.[#], & **Ralf, A.**[#] (2022). Improving the differentiation of closely related males by RMplex analysis of 30 Y-STRs with high mutation rates. *Forensic Science International: Genetics*, *58*, 102682. **(This thesis)**

5. **Ralf, A.**, Zandstra, D., Weiler, N., van Ijcken, W. F., Sijen, T., & Kayser, M. (2021). RMplex: An efficient method for analyzing 30 Y-STRs with high mutation rates. *Forensic Science International: Genetics*, *55*, 102595. **(This thesis)**

6. **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.

7. **Ralf, A.**, Lubach, D., Kousouri, N., Winkler, C., Schulz, I., Roewer, L., ... & Kayser, M. (2020). Identification and characterization of novel rapidly mutating Y-chromosomal short tandem repeat markers. *Human Mutation*, *41*(9), 1680-1696. **(This thesis)**

8. Breslin, K., Wills, B., **Ralf, A.**, Garcia, M. V., Kukla-Bartoszek, M., Pospiech, E., ... & Kayser, M. (2019). HIrisPlex-S system for eye, hair, and skin color prediction from DNA: Massively parallel sequencing solutions for two common forensically used platforms. *Forensic Science International: Genetics*, *43*, 102152.

9. López, C. D., Vidaki, A., **Ralf, A.**, González, D. M., Radjabzadeh, D., Kraaij, R., ... & Kayser, M. (2019). Novel taxonomy-independent deep learning microbiome approach allows for accurate classification of different forensically relevant human epithelial materials. *Forensic Science International: Genetics*, *41*, 72-82.

10. **Ralf, A.**, van Oven, M., González, D. M., de Knijff, P., van der Beek, K., Wootton, S., ... & Kayser, M. (2019). Forensic Y-SNP analysis beyond SNaPshot: High-resolution Y-chromosomal haplogrouping from low quality and quantity DNA using Ion AmpliSeq and targeted massively parallel sequencing. *Forensic Science International: Genetics*, *41*, 93-106.

11. Kayser, M., & **Ralf, A.** (2018). Small number of slowly-mutating (SM) Y-STRs not suitable for forensic and evolutionary applications. *Forensic Science International: Genetics*, *36*, e13.

12. **Ralf, A.**, Montiel González, D., Zhong, K., & Kayser, M. (2018). Yleaf: software for human Y-chromosomal haplogroup inference from next-generation sequencing data. *Molecular Biology and Evolution*, *35*(5), 1291-1294.

13. Adnan, A., Rakha, A., Noor, A., van Oven, M., **Ralf, A.**, & Kayser, M. (2018). Population data of 17 Y-STRs (Yfiler) from Punjabis and Kashmiris of Pakistan. *International Journal of Legal Medicine*, *132*(1), 137-138.

14. Vidaki, A., López, C. D., Carnero-Montoro, E., **Ralf, A.**, Ward, K., Spector, T., ... & Kayser, M. (2017). Epigenetic discrimination of identical twins from blood under the forensic scenario. *Forensic Science International: Genetics*, *31*, 67-80.

15. Adnan, A., **Ralf, A.**, Rakha, A., Kousouri, N., & Kayser, M. (2016). Improving empirical evidence on differentiating closely related men with RM Y-STRs: a comprehensive pedigree study from Pakistan. *Forensic Science International: Genetics*, *25*, 45-51.

16. Chaitanya, L., **Ralf, A.**, van Oven, M., Kupiec, T., Chang, J., Lagacé, R., & Kayser, M. (2015). Simultaneous whole mitochondrial genome sequencing with short overlapping amplicons suitable for degraded DNA using the ion torrent personal genome machine. *Human Mutation*, *36*(12), 1236-1247.

17. Alghafri, R., Goodwin, W., **Ralf, A.**, Kayser, M., & Hadi, S. (2015). A novel multiplex assay for simultaneously analysing 13 rapidly mutating Y-STRs. *Forensic Science International: Genetics*, *17*, 91-98.

18. Zubakov, D., Kokmeijer, I., **Ralf, A.**, Rajagopalan, N., Calandro, L., Wootton, S., ... & Kayser, M. (2015). Towards simultaneous individual and tissue identification: A proof-of-principle study on parallel sequencing of STRs, amelogenin, and mRNAs with the Ion Torrent PGM. *Forensic Science International: Genetics*, *17*, 122-128.

19. Robino, C., **Ralf, A.**, Pasino, S., De Marchi, M. R., Ballantyne, K. N., Barbaro, A., ... & Kayser, M. (2015). Development of an Italian RM Y-STR haplotype database: Results of the 2013 GEFI collaborative exercise. *Forensic Science International: Genetics*, *15*, 56-63.

20. **Ralf, A.**, van Oven, M., Zhong, K., & Kayser, M. (2015). Simultaneous analysis of hundreds of Y-Chromosomal SNP s for high-resolution paternal lineage classification using targeted semiconductor sequencing. *Human Mutation*, *36*(1), 151-159.

21. Ballantyne, K. N.[#], **Ralf, A.**[#], Aboukhalid, R., Achakzai, N. M., Anjos, M. J., Ayub, Q., ... & Kayser, M. (2014). Toward male individualization with rapidly mutating y-chromosomal short tandem repeats. *Human Mutation*, *35*(8), 1021-1032.

22. Liu, F., Hendriks, A. E. J., **Ralf, A.**, Boot, A. M., Benyi, E., Sävendahl, L., ... & Kayser, M. (2014). Common DNA variants predict tall stature in Europeans. *Human Genetics*, *133*(5), 587-597.

23. van Oven, M., Toscani, K., van den Tempel, N., **Ralf, A.**, & Kayser, M. (2013). Multiplex genotyping assays for fine-resolution subtyping of the major human Y-chromosome haplogroups E, G, I, J, and R in anthropological, genealogical, and forensic investigations. *Electrophoresis*, *34*(20-21), 3029-3038.

24. Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., **Ralf, A.**, Kosiniak-Kamysz, A., ... & Kayser, M. (2013). The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Science International: Genetics*, *7*(1), 98-115.

25. Ballantyne, K. N., van Oven, M., **Ralf, A.**, Stoneking, M., Mitchell, R. J., van Oorschot, R. A., & Kayser, M. (2012). MtDNA SNP multiplexes for efficient inference of matrilineal genetic ancestry within Oceania. *Forensic Science International: Genetics*, *6*(4), 425-436.

26. Ballantyne, K. N., Keerl, V., Wollstein, A., Choi, Y., Zuniga, S. B., **Ralf, A.**, ... & Kayser, M. (2012). A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. *Forensic Science International: Genetics*, *6*(2), 208-218.

27. van Oven, M., **Ralf, A.**, & Kayser, M. (2011). An efficient multiplex genotyping approach for detecting the major worldwide human Y-chromosome haplogroups. *International Journal of Legal Medicine*, *125*(6), 879-885.

[#] These authors contributed equally to this work.

PhD Portfolio

Name: Arwin Ralf
Department: Genetic Identification
Promotor: Prof. Dr. M.H. Kayser
Co-promotor: Dr. M.H.D. Larmuseau

Description	ECTs
Courses:	
EMC - Safely working in the Laboratory (2018)	0.3
EMC - CC02A Biostatistical Methods I: Basic Principles Part A (2018)	2
EMC - Scientific Integrity (2018)	0.3
Course in R (2018)	1.8
The data analysis in Python (2019)	1.7
The Biomedical English Writing Course for MSc and PhD students (2019)	2
Genetics course (2019)	3
The Workshop presenting skills for PhD students and Post Docs (2019)	1
Erasmus MC - Writing Successful Grant Proposals (2019)	0.5
Subtotal ECTS	**12.6**
Conferences and seminars:	
Attending and presenting at the 27th ISFG Congress 2017: Seoul (2017)	2
Attending and presenting at Haploid Markers 2018 Conference (2018)	2
Presented at the Gednap (2019)	1
Presented at the ISFG 2019 congress (2019)	2
Seminar at the HID Roadshow 2019 (2019)	1
Presented at the Young Investigators Virtual Summer seminars (2020)	1
Subtotal ECTs	**9**
Teaching activities:	
Supervising MSc student Diego Montiel Gonzalez (2017)	2
Guest Lecture Hanzehogeschool Groningen (2018)	1
Supervising HBO student Delano Lubach (2018)	2
Supervising HBO student Leroy de Visser (2018)	2
Guest lecture on forensic genetics at Hogeschool Zeeland (2019)	1
Supervising MSc student Dion Zandstra (2019)	2
Lecture at the minor Genetica in de Maatschappij (2019)	1
Guest lecture at the Hogeschool Zeeland (2020)	1

Lectures minor Genetics in Society (2020)	2
Supervising Msc student Zehra Koksal (2021)	2
Guest lecture at the Hogeschool Zeeland (2021)	1
Lecture at Symposium Hogeschool Saxion (2021)	1
Lecturer minor Genetics in Society (2021)	1.5
Guest lecture at the Hogeschool Zeeland (2022)	1
Supervision of Bsc student Evelina Parvanona (2022)	2
Subtotal ECTs	**22.5**
Total ECTs	**45.3**

About the author

Arwin Ferdinand Ralf was born in Vlissingen on the 6th of June 1987. He graduated from the Hogeschool Zeeland in 2009, where he had studied Chemistry, with Life Sciences as a major. During his studies he got intrigued by the DNA and wished to learn more about variations between humans and human populations. In 2009, he did his internship at the Erasmus MC in Rotterdam at the Department of Forensic Molecular Biology, performing the final confirmations of the mutations in a large-scale Y-STR mutation rate study. He did his graduation internship at the same department, the subject of his graduation was the usage of bacterial DNA to perform forensic tissue identification. After his graduation he got the opportunity to be employed, starting in 2010, at the department of Forensic Molecular Biology as a research technician, where he worked on a variety of different project mostly focused on human genetics. From 2014 to 2016, the author worked at a Technician Next Generation Sequencing at DDL Diagnostic Laboratory in Rijswijk. There he helped optimizing the next generation workflow and he contributed to the diagnostics analysis of the hepatitis C viruses for customers of the company that were in the process of human clinical trials. In 2016 he returned to the Erasmus MC to his old department led by Prof. Manfred Kayser, which had been renamed to the Department of Genetic Identification, where he was employed as a senior research technician. In 2017, he received special permission from the Doctorate Board of the Erasmus University Rotterdam represented by the rector magnificus: Prof. Dr. H.A.P. Pols, to start the PhD track, despite not having obtained a MSc-degree. Since then until the moment of writing, the author has been working as a senior research technician, while simultaneously working on his PhD research.

Dankwoord / Acknowledgements

Ik draag dit proefschrift op aan mijn grootouders, sinds dit jaar zijn ze allen in het hiernamaals. Mijn opa Emile Ralf is overleden lang voordat ik werd geboren, desalniettemin draag ik zijn Y-chromosoom met mij mee. Mijn opa Eliza de Jong, ik heb zoveel mooie jeugdherinneringen aan u, de tijd die je als zeevarende met je kinderen hebt gemist heb je met je kleinkinderen ingehaald, ik zal die herinneringen altijd bij me dragen. Mijn oma Ilse de Baas, u woonde altijd in Suriname, daarom was onze tijd samen beperkt. Wat me nog heel helder bijstaat is uw bigi yari die ik bij kon wonen in Suriname toen u 85 werd, ik ben blij dat we nog even hebben kunnen dansen. Mijn oma Johanna Dourlein, u heeft het zo lang volgehouden, maar eerder dit jaar was dan ook uw tijd gekomen. Ik vind het vreselijk jammer dat u geen getuige zult zijn van mijn verdediging, het scheelde zo weinig. Maar bovenal ben ik onnoemelijk dankbaar voor al die jaren dat u er nog wel was en alle momenten die we nog wel hebben kunnen delen. U heeft voor mij en mijn gezin zoveel dingen mogelijk gemaakt die anders nooit haalbaar waren geweest. Het verlies is nog vers, als ik dit proefschrift verdedig dan zijn mijn gedachten ook bij u.

Gezien het type onderzoek dat ik doe ben ik me er meer dan gemiddeld van bewust dat zonder elk van mijn grootouders ik er nooit was geweest. Ik draag jullie allen mee in mijn hart, in mijn bloed, in mijn DNA.

Aan mijn ouders:

Pa, onze relatie is niet altijd makkelijk geweest ik vond het lastig dat je zoveel aan het varen was. Ik snap het nu beter, ik ervaar nu zelf de druk die het geeft om een gezin te onderhouden. Als het niet gaat zoals het moet, dan moet het zoals het gaat. Ik ben dankbaar, want door jouw harde werken al die jaren heb ik de kansen gekregen om te komen waar ik nu ben. Ik geloof dat ik jouw gedrevenheid het meegekregen, mijn drang om wezenlijk dingen te op te bouwen. Je bent een goede, lieve, eerlijke en hard wekende man, zo wil ik ook zijn. Ma, je was altijd en bent nog steeds mijn steun en toeverlaat. Als ik het lastig heb en ergens mee zit, dan kan ik altijd bij jou terecht. Ik hoop dat binnenkort het lange zorgen voor jou ook voorbij is zodat je rustig van je oude dag kunt gaan genieten. Wat er ook gebeurt, ik zal er in de toekomst ook altijd voor jou zijn als je me nodig hebt.

Zonder mijn ouders was ik nergens geweest, ik ben jullie oneindig dankbaar voor de zorgeloze en stressvrije jaren die ik bij jullie thuis heb doorgebracht en voor alle hulp die jullie mij daarna nog hebben geboden.

Aan mijn gezin:

Nathalie, bedankt dat je een goede moeder bent voor mijn kinderen. Je hebt je hele volwassen leven alleen maar gezorgd, ik hoop dat je snel ook meer tijd en ruimte zult vinden om voor meer jezelf te zorgen. Er is nog zoveel meer in het leven, ik hoop dat je dat, wat later dan gemiddeld, ook nog allemaal zult gaan ontdekken. Alysha, gup, spannende tijden voor jou, ik ben super trots op wat je al hebt bereikt, maar het begint nu nog maar net. Ik weet dat je een leeftijd hebt waarop je denkt dat je al heel groot bent en dat wij je vooral het leven moeilijk willen maken, maar geloof me, je moeder en ik hebben altijd het beste met je voor. Het leven zit vol plezier en leuke dingen, maar er is ook gevaar, zeker voor een jonge dame. Maak verstandige keuzes, dan blijft je een hoop ellende bespaard. En mocht je dan toch een keer een verkeerde afslag hebben genomen, neem mij dan in vertrouwen, ik zal er altijd voor je zijn om je te helpen en zo nodig uit de problemen te halen. Als je door gaat op het pad dat je nu bewandeld dan ligt vrij letterlijk heel de wereld voor je open, oprecht. Zefyre, mijn oneindig lieve en drukke ventje, je hebt zoveel plezier in mijn leven gebracht. Ik hoop dat je een voor Barcelona of Feyenoord voetballende achtbaanbouwer wordt zoals je nu wilt. Of wat het dan ook zal zijn dat je later wilt worden. Zaiyon, mijn kleine driftige lieve Freek Vonk, ik denk dat jij in je vaders voetstappen gaat treden en iets gaat doen met biologie, maar dan met dieren. Je bent een bijzonder ventje, ik hoop dat je je emoties wat beter leert controleren, want als je lief bent ben je extreem lief, maar als je boos bent dan… Oh ja, jongens, zorgen jullie voor een paar kleinzonen? We moeten ons Y-chromosoom goed conserveren en ik wil over 1000 jaar graag mijn eigen haplogroep.

Net als mijn eigen vader werk ik te hard en werk ik te veel, maar uiteindelijk doe ik dat om mijn gezin een beter leven te geven, zodat iedereen zijn of haar dromen kan najagen, net zoals mijn vader dat voor mij heeft gedaan.

Aan mijn vrienden:

Ik was hier ook nooit gekomen zonder de steun van mijn vrienden, jullie zorgden voor afleiding en dat ik mijn hart kon luchten als het nodig was. Martijn, mijn maatje sinds de brugklas, nu met je eigen mooie gezin, ik ben trots op hoe je je leven helemaal volgens plan hebt ingericht, petje af man. En dan dachten ze op het VWO nog wel dat wij nooit wat in het leven zouden bereiken. Erik, jij bent ook echt een baas, je bent succesvol ontsnapt aan de horeca en dat is niet makkelijk. Je doet het goed man, je werkt aan je eigen ontwikkeling, geniet volop van het leven en staat altijd voor iedereen klaar. Ik heb in mijn leven weinig goedzakken zoals jou ontmoet. Rutger, het heeft je de afgelopen jaren niet meegezeten, net toen je stappen vooruit aan het maken was begon je lichaam je tegen te werken. Maar je hebt een grote stap gezet man, ik wens je heel veel geluk toe

met Rox, je hebt dat verdiend. Ik zou nog veel meer kunnen zeggen over de grote steun die je altijd voor me bent geweest broertje, maar je weet zelf. Nick, altijd goed voor een ongezouten mening. Ondanks je directe benadering en uitgesproken mening ben jij ook gewoon een goedzak. Je timing om te gaan ondernemen had beter gekund, maar als al deze externe factoren een beetje uitdoven dan ga je het daar nog ver mee schoppen. Hopelijk heb je dan ook tijd om wat meer te chillen maat. Dieuwert, je bent ook echt een goede gozer, probeer het leven soms wat zonniger in te zien. Je hebt een baan die dicht staat bij je hobby, en zelfs onder het zware Bax regime presteer je het om stappen vooruit te maken. Ik ga mijn best doen om wat vaker jamsessies met je in te plannen, als we ons best doen kunnen we onze droom verwezenlijken om in de Irish pub te spelen. René, ik ben benieuwd wat jouw volgende stappen worden. Je bent handig en je bent een prettig gezelschap, dus wat mij betreft ligt de wereld voor jou open. Het is mooi dat je de tijd hebt genomen om te genieten van je kleine meid en om je huis op orde te maken. Ilja, ook jouw toekomst ligt nog helemaal open. Ik vind het altijd mooi om te zien hoe je geniet van alle kinderen in de vriendengroep. Je plukt de dag, ik ben wel eens jaloers op hoe je het maximale uit je vrijheid weet te halen. Ga zo door jongen.

Zonder mijn vrienden was het leven maar een saaie bedoeling geweest, dat er nog maar veel bier mag vloeien.

To my (past) colleagues:

First of all I have to express my gratitude to Manfred, you have very literally changed my life by believing in me. In the first years there have been some moments where there were some disagreements, but ultimately all of that led to a very special working relationship where it seems we understand each other very well. Time and time again you have given me the opportunities that I sought, thanks to that I could develop the way I have. I never took the standard road, I have always been trying to do things my way, because of your trust I could do that, with success. I know you are in a rocky phase of your life, I sincerely hope that everything will work out for the best, please know that you can always count on me to repay the support that you have given me.

Kaye, Mannis, Oscar, Fan, Mark and Ying, time flew by, but in my early years in the lab you have all taught me your own valuable lessons. In a way you have all provided me with an unofficial Master's degree, I combined all your valuable lessons together to develop myself into a young scientist. You have all been great mentors, thank you for that.

Iris, Famke, onze tijd samen op het lab was ook heel bijzonder, het was niet alleen gezellig, maar ook op professioneel vlak was je een super goede collega. Ik hoop dat je leert om wat meer voor jezelf op te komen, je bent veel betere en een waardevollere werknemer

dan je zelf misschien denkt. Naast het werk durf ik te zeggen dat we ook echt vrienden waren. Leuke tijden samen met Lisa en Kim en daarna. Mocht je in de toekomst nog eens een keer collega's willen worden, dan ben je altijd welkom.

Eva, terugkijkend was jij de enige die me echt iets heeft geleerd bij DDL, ik denk dat als jij niet op een bepalend moment op vakantie was geweest dat het dan allemaal anders zou zijn gelopen. Ik ben blij voor je dat je een nieuwe stap hebt gemaakt en ik hoop dat je het goed naar je zin hebt bij Naturalis.

Nefeli, grasshopper, you made my return to the lab a lot more enjoyable, actually it was you that inspired me to get back on the track of RM Y-STRs. It was very nice having you as a colleague, you are a very kind person. I am happy for you that you are making big steps in your life by moving back to Greece, as you had planned. Don't be a stranger and let me know if you are back in Rotterdam.

Athina, you are an inspiring person, I could never work as hard as you (nor would I want to), you are probably the most driven person I have ever met. With that attitude you will surely reach to great heights. But I hope you will also find more fulfillment outside of work. Thank you for always being there when I needed advice and to prepare me for what is ahead of me. Whatever the future may hold for you, let's find ways to keep working together.

My fellow PhD students: Celia, Gabriela, Ben and Lucie, you are either finished or almost finished, it was a great pleasure to share this burden with you all. Sometimes I felt like it was not fair because I seemed to have it easier compared to some of you sometimes. It's a rocky road, but the finish line is in sight, I am sure that eventually it will all be worth it and all of us will end up at the places where we are supposed to be.

My students: Kimberley, Delano, Leroy, Faidra, Zehra and Evelina, thank you for putting up with me. All of you seem to have landed on your feet, I am proud of all of you and I hope that every now and then you will remember me and one of the teachings that I passed on to you.

Diego, broodje, we came a long way man. You were vital for me throughout my PhD years, you are a really great guy and I am happy that you seem to have found happiness in Utrecht. Never forget, you are a superstar!

Dion, my protégé, I am looking forward on continuing this journey with you as you work towards your own PhD. Your skill and energy combined with my critical eye and experience are bound to progress the RM Y-STR research far beyond what was achieved in this thesis.

Lastly, two colleagues, although not direct colleagues that I admire: First Sofie Claerhout, sometimes I refer to you as my female Belgian counterpart. It is inspiring how you seem to become a bit of a celebrity in Belgium which is very important. Only with good science communication we can convince the people in power to allow the technologies that we develop to be used in practice. The end goal is always to make the world a safer place, science communication is key and you seem to excel at it. Let's both work hard to make the Y-chromosome the most important chromosome in forensic genetics.

And then of course Walther, you are a role model for me, ever since I first met you at the Haploid Markers meeting in Innsbruck you have made a strong impression on me. It is a great honor and it was my explicit wish that you would take a seat in my Doctoral Committee. I hope you can be a mentor for me in the years to come and help guide me towards the highest office. In any case I hope we will get to meet more frequently than was the case in these past Covid-years.

There are certainly people that I have met along the way that I have not acknowledged personally here. Therefore, I would like to thank everyone wro supported me, worked with me, drank with me and, or believed in me. Someone like me obtaining a PhD shows that there are no limits to what anyone can become.