Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

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] | Manfred Kayser[1]

[1]Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands

[2]Institut für Blutgruppenforschung LGC GmbH, Cologne, Germany

[3]Abteilung für Forensische Genetik, Institut für Rechtsmedizin und Forensische Wissenschaften, Charite´-Universitätsmedizin Berlin, Berlin, Germany

[4]Institut für Rechtsmedizin, Universitätsklinikum Halle, Halle/Saale, Germany

[5]Department of Medical Genetics and Department of Forensic Medicine, Medical University Warsaw, Warsaw, Poland

[6]Department of Forensic Medicine, Wroclaw Medical University, Wroclaw, Poland

[7]Department of Human Genetics, KU Leuven, Leuven, Belgium

[8]Histories VZW, Mechelen, Belgium

Correspondence
Manfred Kayser, Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, 3000 CA Rotterdam, The Netherlands.
Email: m.kayser@erasmusmc.nl

Present address
Nefeli Kousouri, GenomeScan BV, 2333 BZ Leiden, The Netherlands

Christian Winkler, IFB Institut für Blutgruppenforschung GmbH, 50933 Cologne, Germany

Iris Schulz, Abteilung für Forensische Genetik, Institut für Rechtsmedizin, 4056 Basel, Switzerland

Josephine Purps, Berlin Police, Criminal Investigation Department, Forensic Science İnstitute, Berlin, Germany

Funding information
Erasmus MC University Medical Center Rotterdam

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 male relatives. Previously, we identified 13 Y-STR markers with untypically high mutation rates (>10^{-2}), termed rapidly mutating (RM) Y-STRs, and showed that they improved male relative differentiation over standard Y-STRs. By applying a newly developed in silico search approach to the Y-chromosome reference sequence, 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. Their capacity to differentiate males related by 1, 2, and 3 meioses was 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 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.

KEYWORDS
forensic genetics, genetic genealogy, genetic identification, male relative differentiation, mutation rates, rapidly mutating Y-STRs, Y-STRs

1 | 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 (Fregeau & Fourney, 1993; Lygo et al., 1994), 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 (Kayser, 2017), genetic genealogy (Calafell & Larmuseau, 2017), anthropological genetics, and human population history research (Jobling & Tyler-Smith, 2017).

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 with DNA of the male perpetrator's sperm cells (Roewer, 2009). 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 (Gill et al., 2015; Vuichard et al., 2011). 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 (Kayser, 2017). 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 (Kayser, 2017), forensic genealogy (Phillips, 2018), or surname prediction (Claerhout et al., 2020). 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 2,000 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 (Ballantyne et al., 2010). Followed by the first empirical demonstrations of their suitability for male relative differentiation (Ballantyne et al., 2012, 2014), many subsequent studies provided increasing evidence on the value of RM Y-STRs for differentiating related, including closely related, and also unrelated men (Adnan, Ralf, Rakha, Kousouri, & Kayser, 2016; Alghafri, Goodwin, & Hadi, 2013; Boattini et al., 2016, 2019; Lang et al., 2017; Niederstätter, Berger, Kayser, & Parson, 2016; Robino et al., 2015; Salvador et al., 2019; Turrina, Caratti, Ferrian, & De Leo, 2016; Westen et al., 2015; Zgonjanin, Alghafri, Antov et al., 2017). In genetic genealogy too, RM Y-STRs are advantageous as they provide improved

differentiation of unrelated individuals (Ballantyne et al., 2014) and they allow distinguishing closely related from more distantly related males by taking the number of observed mutations into account (Larmuseau et al., 2019).

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 (Roewer, 2019). 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 (Adnan et al., 2016), 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 (Ballantyne et al., 2010; Goedbloed et al., 2009); 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 (Willems et al., 2016), or deep-rooted male pedigrees (Boattini et al., 2019; Claerhout et al., 2018) 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 with 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; Willems et al., 2016). 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, that is, analyzing 1,000 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, which can lead to inaccurate estimations of the mutation rates (Larmuseau et al., 2013; Willems et al., 2016). Another disadvantage of both of these approaches is the potential presence of parallel mutations, hidden mutations and multistep mutations, which all could lead to increased error in the mutation rate estimates obtained (Claerhout, Van der Haegen, Vangeel, Larmuseau, & Decorte, 2019). 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 vs. repeat loss) and quantifying the step-wise nature of the repeat mutations (single step vs. multistep) unambiguously.

Since our previous Y-STR mutation study (Ballantyne et al., 2010) 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.

2 | MATERIALS AND METHODS

2.1 | Editorial policies and ethical considerations

The biological material, from which the DNA samples used in this study were previously extracted, had been collected by the respective coauthors for paternity testing purposes with the donors' given agreement that left-over materials can be used for genetic research. These DNA samples were fully anonymized (i.e., the key to link the samples, and the data produced from the samples, with the sample donors was destroyed). The only information of the sample donors that was kept together with the DNA sample, and used in this study, was father–son relationship, age and place of sample collection, which does not reflect personal data (i.e., data that can be linked to an identified or identifiable person). Also the use of these DNA samples in this study for investigating Y-STR mutability does not produce any personal data under this widely accepted definition. Because this study does not use nor produce personal data, it is outside of the remit of national or international (such as European Union) privacy protection laws. As far as other research ethics aspects beyond privacy protection are concerned, sample collection took place at a time when no formal ethical board approval was possible for this based on national regulations in place in the respective countries at the time, and in principle cannot be obtained retrospectively. These DNA samples had been used for the same purpose of investigating Y-STR mutability in two previous publications (Ballantyne et al., 2010; Goedbloed et al., 2009).

2.2 | 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 catalog containing all Y-STRs present in the latest assembly of the human genome (GRCh38), by using the publically available software Tandem repeats finder (Benson, 1999). The following parameters were set in the software: Match = 2, Mismatch = 100, Delta = 100, PM = 80, PI = 10, Minscore = 12, and MaxPeriod = 5. These settings resulted in a catalog 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 catalog we discarded all repeats with a motif size < 3, as such markers suffer from too much stutter (Hauge & Litt, 1993). Y-STRs located in pseudoautosomal regions were also excluded, because such regions do not contain male-specific loci (Mensah et al., 2014; Poriswanish et al., 2018). 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 catalog, 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, that is, <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 multicopy 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 (Ballantyne et al., 2010; Brinkmann, Klintschar, Neuhuber, Hühne, & Rolf, 1998; Eckert & Hile, 2009; Ellegren, 2004; Kayser et al., 2000, 2004; Kelkar, Tyekucheva, Chiaromonte, & Makova, 2008; Willems et al., 2016): (a) the length (i.e., number of repeats) of the uninterrupted repeat stretches, (b) the number of repeat stretches in a sequence, (c) the marker being a single-copy, or a multicopy marker, and (d) 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 (Ballantyne et al., 2010; Brinkmann et al., 1998; Eckert & Hile, 2009; Ellegren, 2004; Kayser et al., 2000, 2004; Kelkar et al., 2008).

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 RM 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 multicopy 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 (Ballantyne

DYS627

CTAGGTGACAGCGCAGGATTCCATCTAAAAACAAAAATCAAAGAGAGACACAGGGAGAAAGAAAGAAAGGAAAAAAGAAAGAA
AGAGAGAGAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAA
GAAAGAAAAAAGAGAAAGAGAGAGGGGAGGGGAGGAGGGGAGGGGAGGAAGTAAAGAAGGTAAGAAGGAAGGATGGAAGTAAGGAAG
GAAAGAAGGAAGGAAGGAACGAAGGAGAGAATGAAAGGCATAAAACCAAAGGGAAATGAATAAAAATAAAAACTTGCCATTTG
CTCATTATCC

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

DYS526b

TCAGCAAACACCTTAGCTAATCATGATTCTTTACCTTCTGGTGAACTGATCCAAACCTTTACTTATTTTCTTCCTTTTTTCCTCTCTCC
TTCCCTCCCTCCCTCTTCTTCCTTTCTTTCTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCC
TTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCTTTCCTTCTTTCCTTCATTTCTTCTTTCTTTCTCTTTACCTTCTGGGAAACTG
ATCCAAACCTTTACTTACTTTCTATTTTTCTCTATTTTTTCTTTCTCTATTTTTTTCTCTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCC
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). RM rapidly mutating; STRs, short tandem repeats

et al., 2010; Brinkmann et al., 1998; Eckert & Hile, 2009; Kelkar et al., 2008), therefore an exponential function was derived empirically from the Y-STRs and mutation rates described previously (Ballantyne et al., 2010). 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 (Ballantyne et al., 2010) contains two repeat stretches, one of six and one of 18 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, that is, trinucleotide, or pentanucleotide repeats, when considering similar numbers of repeat units (Ballantyne et al., 2010; Eckert & Hile, 2009). 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, that is, nomenclature, genomic locations, and so forth were stored in the Human Genome Database, which, however, is no longer available. 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. In addition, 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 multicopy markers. We used numbers larger than one thousand since such numbers had not yet been used to describe Y-STRs.

2.3 | 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, polymerase chain reaction (PCR) primer design was performed using Bisearch (Tusnady, Simon, Varadi, & Aranyi, 2005) 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, that is, 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 with singleplex PCR) genotyping of the large number of DNA samples from fathers and their sons we considered in this study. Autodimer software (Vallone & Butler, 2004) 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, that is, primer sequences and mutability prediction scores, of the cRM Y-STRs can be found in Table S1. Table S1 also shows repeat descriptions based on the HGVS nomenclature system (den Dunnen et al., 2016). However, in this study we did not sequence the markers and, therefore, we lack knowledge about the sequence variability, hence the repeat descriptions are done solely based on the GRCh38 reference sequence. 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.), oligonucleotides 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) and with 0.3 μl of ILS600 size standard (Promega Corporation). This mixture was incubated at 95°C for 3 min and rapidly cooled on ice for 5 min. CE was performed on an ABI3130XL Genetic Analyzer (Thermo Fisher Scientific) using sixteen 36 cm capillaries and POP-7 Polymer (Thermo Fisher Scientific). 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).

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 (Ballantyne et al., 2010), 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 (Ballantyne et al., 2010). 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 multicopy markers it was decided that peak height ratio differences would not be interpreted

as mutations, for example, a hypothetical multicopy 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, for example, 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 multicopy markers that is, call both the father and son as 15–16 in the example given above.

2.4 | Mutational data analysis

Statistical data analyses were performed using R version 3.6.2 (R Core Team, 2013; https://www.r-project.org) in Rstudio Version 1.2.5033 (RStudio Team, 2015; https://rstudio.com). Unless stated otherwise functions standardly imbedded in R were used.

2.4.1 | Validation of mutability prediction score

To validate whether the mutability prediction score was a suitable predictor for Y-STR mutation rate, a linear regression analysis was performed to show the correlation between the mutation rates and the mutability prediction score of 185 Y-STRs from our previous mutation rate study (Ballantyne et al., 2010). In addition, these 185 Y-STRs 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 $< 5.0 \times 10^{-3}$ (MM Y-STRs); fastly mutating: $n = 19$ mutation rates $\geq 5.0 \times 10^{-3}$ and $< 10^{-2}$ (FM Y-STRs); and RM Y-STRs: $n = 14$ mutation rates $\geq 10^{-2}$ (RM Y-STRs). Note that the A and B parts of the multicopy RM Y-STR marker DYF403S1 were considered separately in these analyses, 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.

2.4.2 | Mutation rate estimation

Mutation rates were calculated in a locus-specific manner using the frequentist approach that is, 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 multicopy 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 (Fay, 2010).

2.4.3 | 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:

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

2.4.4 | 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, that is, 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.

2.4.5 | 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 < 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 the Wilcoxon rank sum test and with Bonferroni p-value adjustments in RStudio.

2.4.6 | 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 (Ballantyne et al., 2010), 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.

3 | RESULTS AND DISCUSSION

3.1 | 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 (Ballantyne et al., 2010). A statistically significant positive correlation was observed with an R^2 of .53 ($p < 2.2 \times 10^{-16}$). However, a limitation of the used data set is that it contains many markers (51% of total Y-STRs analyzed) with either just a single, or no mutation observed in the nearly 2,000 father–son pairs analyzed in the previous study. This makes the mutation rates estimated for such markers less reliable (Willems et al., 2016) 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 data set to visualize the differences in mutability prediction scores between Y-STR markers using four marker groups defined by their mutation rates: SM Y-STRs, MM Y-STRs, FM Y-STRs, and 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 with all other three groups MM Y-STRs, FM Y-STRs, and RM Y-STRs (p-values of 1.7×10^{-7}, 3.6×10^{-7}, and 1.7×10^{-8}, respectively). MM Y-STRs showed significant p-values compared with FM Y-STRs and RM Y-STRs (p-values of .0092 and 7.2×10^{-8}, respectively). Comparing FM Y-STRs with RM Y-STRs resulted in a significant p-value of .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
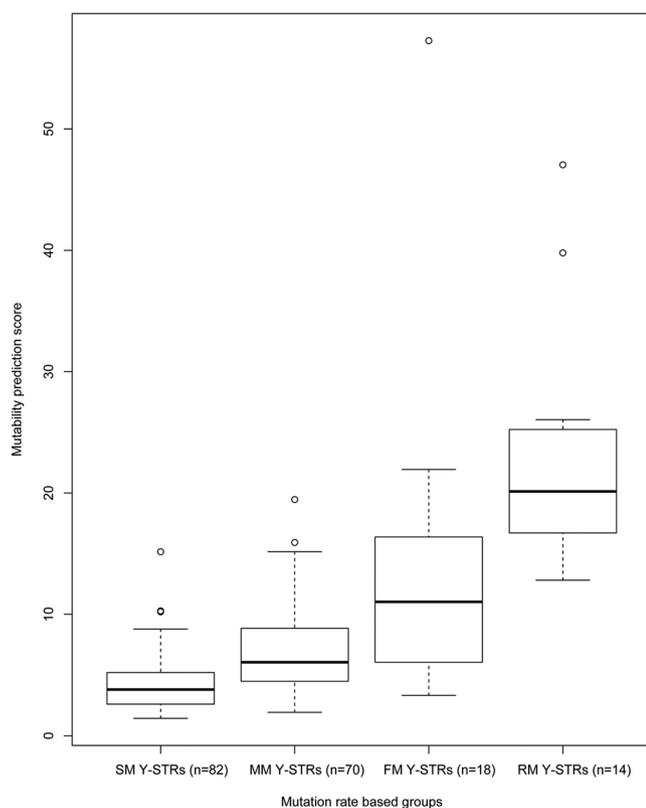
FIGURE 2 Boxplots showing the distributions of the newly developed mutability prediction scores among four groups of Y-STR markers as defined by mutation rate: (a) slowly mutating (SM) Y-STRs (mutation rate < 10^{-3}), (b) moderately mutating (MM) Y-STRs (mutation rate ≥ 10^{-3} and <5×10^{-3}), (c) fast mutating (FM) Y-STRs (mutation rate ≥ 5×10^{-3} and <10^{-2}), and (d) rapidly mutating (RM) Y-STRs (mutation rate ≥ 10^{-2}) based on Y-STRs and their mutation rate estimates from Ballantyne et al. (2010). STRs, short tandem repeats

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 (Ballantyne et al., 2010) were found among the top candidates (before taking multiple repeat stretches and multicopy 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.

3.2 | 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 2 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 .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 (Ballantyne et al., 2010). For seven mutations in our present study, the direction could not be unambiguously assigned due to the multicopy 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 .012 and .077, respectively), while in DYS1006 and DYS1017 it were predominantly repeat expansions with 78% and 77%, respectively (p-values .180 and .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 (Ballantyne et al., 2010). In the present study, only 3% of the observed mutations were two-step mutations and <1% were three-step mutations (Table 1). Notably, our present data set 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.

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

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 (×10⁻³)	95% Confidence interval (×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

(Continues)

TABLE 1 (Continued)

Name	No. of father–son pairs genotyped	No. of mutations observed	Mutation rate (×10⁻³)	95% Confidence interval (×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.

novel Y-STRs not previously described at all, and four Y-STRs previously described in population studies. The previously discovered Y-STRs were: DYS713 (Leat, Ehrenreich, Benjeddou, Cloete, & Davison, 2007), later also described as DYS685 (Maybruck, Hanson, Ballantyne, Budowle, & Fuerst, 2009); DYS711 (Leat et al., 2007), later also described as DYS688 (Maybruck et al., 2009); DYS712 (Leat et al., 2007); and CDY (included in commercial products of FamilyTreeDNA), later also described as DYS724 (Jacobs et al., 2009). Three of those markers had only population data and no mutation data previously reported: DYS711 (Leat et al., 2007; Maybruck et al., 2009; Zhang, Yang, Niu, & Guo, 2012); DYS712; DYS713 (Leat et al., 2007; Liu et al., 2019; Maybruck et al., 2009; Zhang et al., 2012). For one of the previously discovered Y-STR markers, DYS724, mutation data were previously inferred from population data (Chandler, 2006) and later from deep-rooted pedigrees (Boattini et al., 2019; Claerhout et al., 2018), 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 (Leat et al., 2007), and population data were published: DYS714 (Leat et al., 2007; Liu et al., 2019; Zhang et al., 2012). 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 (Hanson & Ballantyne, 2006), 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 (Ballantyne et al., 2010) 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, that is, 7% RM Y-STRs and 9% FM Y-STRs identified in our previous unbiased screening study, including 186 Y-STRs (Ballantyne et al., 2010). 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 with the unbiased, massive screening approach applied previously (Ballantyne et al., 2010). 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 2.6×10^{-2}, which is higher compared with that of the set of previously identified 13 RM Y-STRs with 1.6×10^{-2} (Adnan et al., 2016). 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 6.9×10^{-2} (Adnan et al., 2016). In comparison, the most mutable novel RM Y-STR identified in the present study, DYF1001, has a slightly lower estimated mutation rate of 5.2×10^{-2}. When combining the 12 novel with the 13 previous RM Y-STRs and ranking them according to their empirically derived mutation rate estimates with Rank 1 going to the marker with the highest mutation rate, Rank 2–6 go to 5 of the 12 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 multicopy markers (five) compared with 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 10 out of the 27 cRM Y-STRs were multicopy markers. Of these 10, 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 (Ballantyne et al., 2010).

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, for example, 9.9×10^{-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, for example, all 10 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 data sets. 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 (Willems et al., 2016). In the future, accurate third generation sequencing technologies like Pacbio's single molecule, real-time sequencing may help to overcome these limitations. The future analysis of high-quality, high-coverage, and long read whole genome sequences (Vollger et al., 2019) 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.

3.3 | 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 one 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 1–10 meioses, and compared it with the estimates calculated in the same way for the 13 previously identified RM Y-STRs (Ballantyne et al., 2010). 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 with 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 (Kayser, 2017) and other fields such as genetic genealogy (Calafell & Larmuseau, 2017).

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 (Adnan et al., 2016). 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 (Adnan et al., 2016). 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.

3.4 | Internal and external factors influencing mutability

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

It is generally accepted that the length of an STR repeat, that is, the number of repeats, is the most predominant driving factor of STR including Y-STR mutability (Ballantyne et al., 2010; Brinkmann et al., 1998; Eckert & Hile, 2009; Ellegren, 2004; Kayser et al., 2000; Kelkar et al., 2008; Willems et al., 2016). 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 with 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
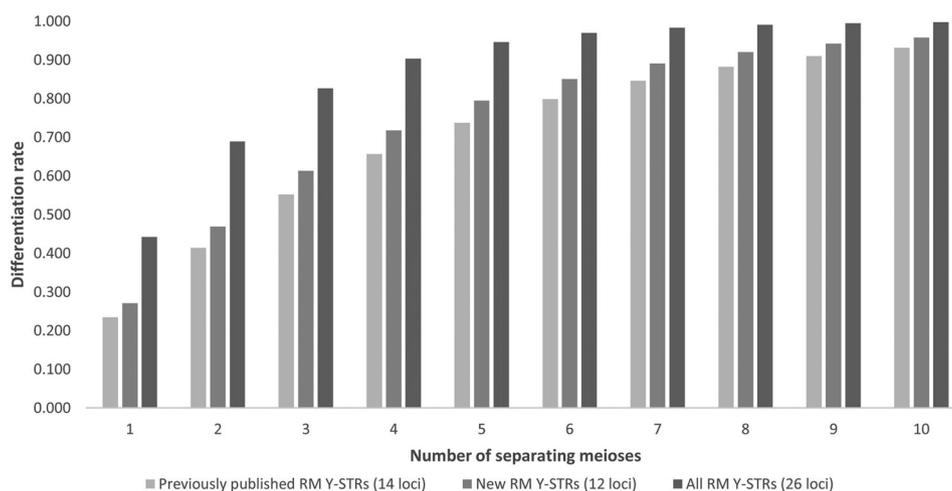
FIGURE 3 Male relative differentiation capacities calculated from the respective locus-specific mutation rate estimates for (a) the 13 previously established RM Y-STRs (Ballantyne et al., 2010), DYF403S1a and DYF403S1b were considered making a total of 14 loci. (b) The 12 novel RM Y-STRs identified in the present study, and (c) the combined set of 25 currently known RM Y-STRs, for male relative pairs separated by 1–10 meioses, respectively. RM, rapidly mutating; STRs, short tandem repeats
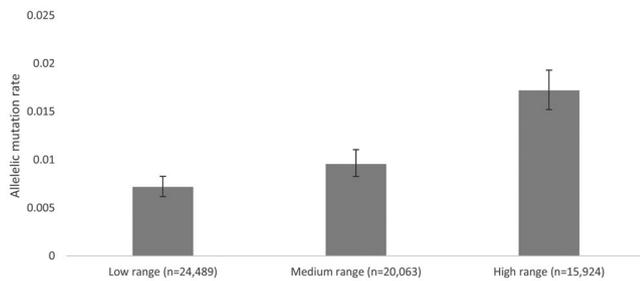
FIGURE 4 Y-STR allelic mutation rates (the number of mutations per allele per meiosis) of the genotyped 1,616 fathers according to the (a) low, (b) medium, and (c) high range allele groups (tertiles) as defined by the father's allelic fragment length based on the 27 cRM Y-STRs highlighted by our in silico approach. cRM, candidate rapidly mutating; STRs, short tandem repeats


FIGURE 5 Y-STR repeat mutation expansion and contraction proportions according to the (a) low, (b) medium, and (c) 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% confidentiality interval. cRM, candidate rapidly mutating; STRs, short tandem repeats

found between the low and medium allele ranges, with an adjusted p-value of .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}.

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 (Claerhout et al., 2018). 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 (Kayser et al., 2005). 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 (Adnan et al., 2016; Ballantyne et al., 2014; Boattini et al., 2016; Lang et al., 2017; Zgonjanin, Alghafri, Almheiri et al., 2017). 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.

3.4.2 | 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 the 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
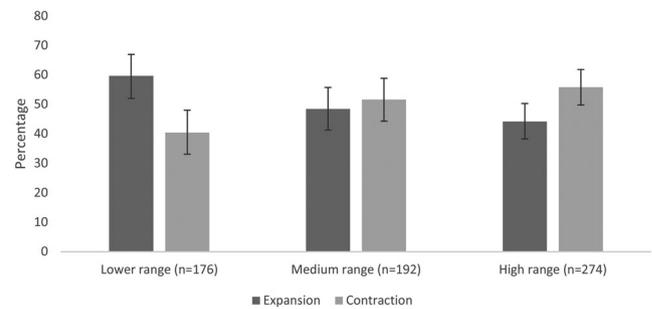
contraction (p-value .012), and a low, yet nonsignificant difference in the high allele range, with more contractions than expansions (p-value .061). In the medium allele range, however, the expansions and contractions appeared to be more balanced, as is also reflected in a nonsignificant p-value of .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 (Ballantyne et al., 2010). 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 (Xu, Peng, Fang, & Xu, 2000).

3.4.3 | 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 (Ballantyne et al., 2010; Claerhout et al., 2018; Gusmao et al., 2005; Kong et al., 2012; Sun et al., 2012). However, other studies reported no such, or only a small effect (Dupuy, Stenersen, Egeland, & Olaisen, 2004; Forster et al., 2015), 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
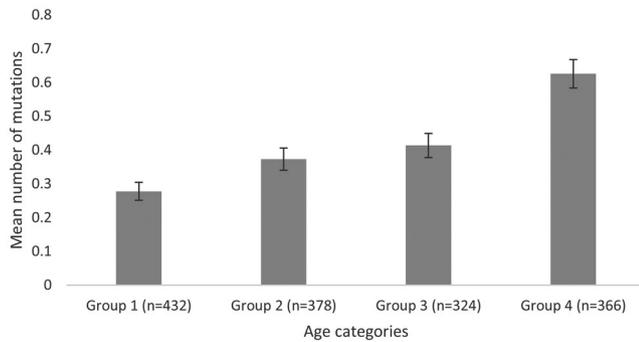
FIGURE 6 Mean number of observed Y-STR mutations according to four categories defined by the father's age at time of conception of his son, the age groups were defined as the quartiles. Group 1: 15–23 years old, Group 2: 24–29 years old, Group 3: 30–36 years old, and Group 4: 37–66 years old, based on 27 cRM Y-STRs highlighted by our in silico approach. cRM, candidate rapidly mutating; STRs, short tandem repeats

impact on the number of observed mutations in our study. In the oldest age group there was a more than a two-fold increase in the mean number of Y-STR mutations observed compared with the youngest age group. A pairwise comparisons using the 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 .0018 compared with Group 1, 2, and 3, respectively). In addition, the second oldest age Group 3 showed significantly more Y-STR mutations than the youngest age Group 1 (p-value of .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 (Ballantyne et al., 2010; Brinkmann et al., 1998; Claerhout et al., 2018; Gusmao et al., 2005; Sun et al., 2012). 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 (Khandwala, Zhang, Lu, & Eisenberg, 2017), there also are strong differences between populations based on various reasons including cultural and economic factors (Young Jr, 2011) that shall be considered for the data interpretation in future studies.

3.4.4 | 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 (Eckert & Hile, 2009; Kelkar et al., 2008), while others did not see such (Ballantyne

et al., 2010). 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, for example, 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 multicopy 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 (Tables 2 and 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-STRs) 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.

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

TABLE 2 Differences in observed STR sequence motifs between RM Y-STRs and non-RM Y-STRs

Motif	RM Y-STRs[a]	Non-RM Y-STRs[b]	p-value
[AAAG]	33 in 34	19 in 117	**<.0001**
[AAGG]	12 in 34	22 in 117	.0606
[AGAT]	1 in 34	37 in 117	**.0003**
[AAAT]	1 in 34	37 in 117	**.0003**
[AAG]	6 in 6	8 in 60	**<.0001**
[AGG]	3 in 6	3 in 60	**.0078**
[AAT]	0 in 6	34 in 60	**.0100**
[AAC]	0 in 6	15 in 60	.3234

Note: Significant p-values (Fisher's exact test) are shown in bold.
Abbreviations: RM, rapidly mutating; STRs, short tandem repeats.
[a]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.
[b]These represent non-RM Y-STRs (mutation rate $< 10^{-2}$ mutations per marker per meiosis) from a previous study (Ballantyne et al., 2010).

repeats respectively (Table S2), which is considerably lower than we found for the RM Y-STRs (p < .0001 and .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 (Tables 2 and 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 .0003 in both cases). Similarly, among the 60 trinucleotide non-RM Y-STR loci from Ballantyne et al. (2010), the AAG and AGG sequence motif families were found only in 13% and 5%, respectively (p < .0001 and .0078, respectively, Tables 2 and S2), while their most frequently observed motifs were AAT and AAC at 57% and 25%, respectively (Tables 2 and S2), which were completely absent in the six trinucleotide RM Y-STR loci (p-values .0100 and .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, that is, 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 (Eckert & Hile, 2009; Kelkar et al., 2008). 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, and AAGG; Eckert & Hile, 2009; Slebos, Oh, Umbach, & Taylor, 2002; Zhao, Bacolla, Wang, & Vasquez, 2010). Whether, this would affect the mutability directly, or rather impacts on the direction of mutations (Shah, Hile, & Eckert, 2010), 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).

4 | CONCLUSIONS

We developed and herewith 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 this 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 with 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 case work and genealogical studies.

CONFLICT OF INTERESTS

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

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.

ORCID

Arwin Ralf http://orcid.org/0000-0001-6052-0807
Lutz Roewer http://orcid.org/0000-0001-9383-4941
Rafal Ploski http://orcid.org/0000-0001-6286-5526
Maarten H. D. Larmuseau http://orcid.org/0000-0002-5974-7235
Manfred Kayser http://orcid.org/0000-0002-4958-847X

REFERENCES

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. https://doi.org/10.1016/j.fsigen.2016.07.005

Alghafri, R., Goodwin, W., & Hadi, S. (2013). Rapidly mutating Y-STRs multiplex genotyping panel to investigate UAE population. *Forensic Science International: Genetics Supplement Series*, 4(1), e200–e201. https://doi.org/10.1016/j.fsigss.2013.10.103

Ballantyne, K. N., Goedbloed, M., Fang, R., Schaap, O., Lao, O., Wollstein, A., & Brauer, S. (2010). Mutability of Y-chromosomal microsatellites: Rates, characteristics, molecular bases, and forensic implications. *The American Journal of Human Genetics*, 87(3), 341–353. https://doi.org/10.1016/j.ajhg.2010.08.006

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. https://doi.org/10.1016/j.fsigen.2011.04.017

Ballantyne, K. N., Ralf, A., Aboukhalid, R., Achakzai, N. M., Anjos, M. J., Ayub, Q., & Berger, B. (2014). Toward male individualization with

rapidly mutating Y-chromosomal short tandem repeats. *Human Mutation*, 35(8), 1021–1032. https://doi.org/10.1002/humu.22599

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, 27(2), 573–580. https://doi.org/10.1093/nar/27.2.573

Boattini, A., Sarno, S., Bini, C., Pesci, V., Barbieri, C., De Fanti, S., & Ferri, G. (2016). Mutation rates and discriminating power for 13 rapidly-mutating Y-STRs between related and unrelated individuals. *PloS One*, 11(11), e0165678. https://doi.org/10.1371/journal.pone.0165678

Boattini, A., Sarno, S., Mazzarisi, A. M., Viroli, C., De Fanti, S., Bini, C., & Luiselli, D. (2019). Estimating Y-STR mutation rates and TMRCA through deep-rooting Italian pedigrees. *Scientific Reports*, 9(1), 9032. https://doi.org/10.1038/s41598-019-45398-3

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

Calafell, F., & Larmuseau, M. H. D. (2017). The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research. *Human Genetics*, 136(5), 559–573. https://doi.org/10.1007/s00439-016-1740-0

Chandler, J. F. (2006). Estimating per-locus mutation rates. *Journal of Genetic Genealogy*, 2, 27–33.

Claerhout, S., Roelens, J., Van der Haegen, M., Verstraete, P., Larmuseau, M. H. D., & Decorte, R. (2020). Ysurnames? The patrilineal Y-chromosome and surname correlation for DNA kinship research. *Forensic Science International: Genetics*, 44, 102204. https://doi.org/10.1016/j.fsigen.2019.102204

Claerhout, S., Van der Haegen, M., Vangeel, L., Larmuseau, M. H. D., & Decorte, R. (2019). A game of hide and seq: Identification of parallel Y-STR evolution in deep-rooting pedigrees. *European Journal of Human Genetics*, 27(4), 637.

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

den Dunnen, J. T., Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., McGowan-Jordan, J., & Taschner, P. E. M. (2016). HGVS recommendations for the description of sequence variants: 2016 update. *Human Mutation*, 37(6), 564–569. https://doi.org/10.1002/humu.22981

Dupuy, B. M., Stenersen, M., Egeland, T., & Olaisen, B. (2004). Y-chromosomal microsatellite mutation rates: Differences in mutation rate between and within loci. *Human Mutation*, 23(2), 117–124. https://doi.org/10.1002/humu.10294

Eckert, K. A., & Hile, S. E. (2009). Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. *Molecular Carcinogenesis*, 48(4), 379–388. https://doi.org/10.1002/mc.20499

Ellegren, H. (2004). Microsatellites: Simple sequences with complex evolution. *Nature Reviews Genetics*, 5(6), 435. https://doi.org/10.1038/nrg1348

Fay, M. P. (2010). Two-sided exact tests and matching confidence intervals for discrete data. *R Journal*, 2(1), 53–58.

Forster, P., Hohoff, C., Dunkelmann, B., Schürenkamp, M., Pfeiffer, H., Neuhuber, F., & Brinkmann, B. (2015). Elevated germline mutation rate in teenage fathers. *Proceedings of the Royal Society B: Biological Sciences*, 282(1803), 20142898. https://doi.org/10.1098/rspb.2014.2898

Fregeau, C. J., & Fourney, R. M. (1993). DNA typing with fluorescently tagged short tandem repeats: A sensitive and accurate approach to human identification. *Biotechniques*, 15(1), 100–119.

Gill, P., Haned, H., Bleka, O., Hansson, O., Dørum, G., & Egeland, T. (2015). Genotyping and interpretation of STR-DNA: Low-template, mixtures and database matches—twenty years of research and development. *Forensic Science International: Genetics*, 18, 100–117. https://doi.org/10.1016/j.fsigen.2015.03.014

Goedbloed, M., Vermeulen, M., Fang, R. N., Lembring, M., Wollstein, A., Ballantyne, K., & Roewer, L. (2009). Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFlSTR® Yfiler® PCR amplification kit. *International Journal of Legal Medicine*, 123(6), 471. https://doi.org/10.1007/s00414-009-0342-y

Gusmao, L., Sánchez-Diz, P., Calafell, F., Martin, P., Alonso, C. A., Alvarez-Fernandez, F., & Bravo, M. L. (2005). Mutation rates at Y chromosome specific microsatellites. *Human Mutation*, 26(6), 520–528. https://doi.org/10.1002/humu.20254

Hanson, E. K., & Ballantyne, J. (2006). Comprehensive annotated STR physical map of the human Y chromosome: Forensic implications. *Legal Medicine*, 8(2), 110–120. https://doi.org/10.1016/j.legalmed.2005.10.001

Hauge, X. Y., & Litt, M. (1993). A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR. *Human Molecular Genetics*, 2(4), 411–415. https://doi.org/10.1093/hmg/2.4.411

Jacobs, M., Janssen, L., Vanderheyden, N., Bekaert, B., Van de Voorde, W., & Decorte, R. (2009). Development and evaluation of multiplex Y-STR assays for application in molecular genealogy. *Forensic Science International: Genetics Supplement Series*, 2(1), 57–59. https://doi.org/10.1016/j.fsigss.2009.08.114

Jobling, M. A., & Tyler-Smith, C. (2017). Human Y-chromosome variation in the genome-sequencing era. *Nature Reviews Genetics*, 18(8), 485. https://doi.org/10.1038/nrg.2017.36

Kayser, M. (2017). Forensic use of Y-chromosome DNA: A general overview. *Human Genetics*, 136(5), 621–635. https://doi.org/10.1007/s00439-017-1776-9

Kayser, M., Kittler, R., Erler, A., Hedman, M., Lee, A. C., Mohyuddin, A., & Jobling, M. A. (2004). A comprehensive survey of human Y-chromosomal microsatellites. *The American Journal of Human Genetics*, 74(6), 1183–1197. https://doi.org/10.1086/421531

Kayser, M., Lao, O., Anslinger, K., Augustin, C., Bargel, G., Edelmann, J., & Henke, L. (2005). Significant genetic differentiation between Poland and Germany follows present-day political borders, as revealed by Y-chromosome analysis. *Human Genetics*, 117(5), 428–443. https://doi.org/10.1007/s00439-005-1333-9

Kayser, M., Roewer, L., Hedman, M., Henke, L., Henke, J., Brauer, S., & Dobosz, T. (2000). 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*, 66(5), 1580–1588. https://doi.org/10.1086/302905

Kelkar, Y. D., Tyekucheva, S., Chiaromonte, F., & Makova, K. D. (2008). The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Research*, 18(1), 30–38. https://doi.org/10.1101/gr.7113408

Khandwala, Y. S., Zhang, C. A., Lu, Y., & Eisenberg, M. L. (2017). The age of fathers in the USA is rising: An analysis of 168 867 480 births from 1972 to 2015. *Human Reproduction*, 32(10), 2110–2116. https://doi.org/10.1093/humrep/dex267

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., & Jonasdottir, A. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488(7412), 471. https://doi.org/10.1038/nature11396

Lang, M., Ye, Y., Li, J., Zhang, Y., Yuan, T., & Hou, Y. (2017). Comprehensive mutation analysis of 53 Y-STR markers in father-son pairs. *Forensic*

Science International: Genetics Supplement Series, 6, e57–e58. https://doi.org/10.1016/j.fsigss.2017.09.004

Larmuseau, M. H. D., van den Berg, P., Claerhout, S., Calafell, F., Boattini, A., Gruyters, L., & Wenseleers, T. (2019). A historical-genetic reconstruction of human extra-pair paternity. *Current Biology*, 29(23), 4102–4107. https://doi.org/10.1016/j.cub.2019.09.075. e4107.

Larmuseau, M. H. D., Vanoverbeke, J., Van Geystelen, A., Defraene, G., Vanderheyden, N., Matthys, K., & Decorte, R. (2013). 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*, 280(1772), 20132400.

Leat, N., Ehrenreich, L., Benjeddou, M., Cloete, K., & Davison, S. (2007). Properties of novel and widely studied Y-STR loci in three South African populations. *Forensic Science International*, 168(2-3), 154–161. https://doi.org/10.1016/j.forsciint.2006.07.009

Liu, J., Wang, R., Hao, T., Cheng, X., Guo, J., Yun, K., & Zhang, G. (2019). 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*, 7(1), 95–97. https://doi.org/10.1016/j.fsigss.2019.09.037

Lygo, J. E., Johnson, P. E., Holdaway, D. J., Woodroffe, S., Kimpton, C. P., Gill, P., & Clayton, T. M. (1994). The validation of short tandem repeat (STR) loci for use in forensic casework. *International Journal of Legal Medicine*, 107(2), 77–89. https://doi.org/10.1007/BF01225493

Maybruck, J. L., Hanson, E., Ballantyne, J., Budowle, B., & Fuerst, P. A. (2009). 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), 11–20. https://doi.org/10.1016/j.fsigen.2009.03.004

Mensah, M. A., Hestand, M. S., Larmuseau, M. H. D., Isrie, M., Vanderheyden, N., Declercq, M., & Van Esch, H. (2014). Pseudoautosomal region 1 length polymorphism in the human population. *PLoS Genetics*, 10(11), e1004578. https://doi.org/10.1371/journal.pgen.1004578

Niederstätter, H., Berger, B., Kayser, M., & Parson, W. (2016). 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*, 24, 180–193. https://doi.org/10.1016/j.fsigen.2016.07.009

Phillips, C. (2018). The Golden State Killer investigation and the nascent field of forensic genealogy. *Forensic Science International: Genetics*, 36, 186–188. https://doi.org/10.1016/j.fsigen.2018.07.010

Poriswanish, N., Neumann, R., Wetton, J. H., Wagstaff, J., Larmuseau, M. H. D., Jobling, M. A., & May, C. A. (2018). Recombination hotspots in an extended human pseudoautosomal domain predicted from double-strand break maps and characterized by sperm-based crossover analysis. *PLoS Genetics*, 14(10), e1007680. https://doi.org/10.1371/journal.pgen.1007680

R Core Team. (2013). *R: A language and environment for statistical computing.*

Robino, C., Ralf, A., Pasino, S., De Marchi, M. R., Ballantyne, K. N., Barbaro, A., & Di Gaetano, C. (2015). Development of an Italian RM Y-STR haplotype database: Results of the 2013 GEFI collaborative exercise. *Forensic Science International: Genetics*, 15, 56–63. https://doi.org/10.1016/j.fsigen.2014.10.008

Roewer, L. (2009). Y chromosome STR typing in crime casework. *Forensic Science, Medicine, and Pathology*, 5(2), 77–84. https://doi.org/10.1007/s12024-009-9089-5

Roewer, L. (2019). Y-chromosome short tandem repeats in forensics—Sexing, profiling, and matching male DNA. *Wiley Interdisciplinary Reviews: Forensic Science*, 1(4), e1336. https://doi.org/10.1002/wfs2.1336

RStudio Team (2015). *RStudio: integrated development for R.* RStudio, Inc., Boston, MA. Retrieved from http://www.rstudio.com

Salvador, J. M., Rodriguez, J. J. R. B., Carandang, L. C. D. L., Agmata, A. B., Honrado, M. L. D., Delfin, F. C., & De Ungria, M. C. A. (2019). Filipino DNA variation at 36 Y-chromosomal short tandem repeat (STR) marker units. *Philippine Journal of Science*, 148, 43–52.

Shah, S. N., Hile, S. E., & Eckert, K. A. (2010). Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. *Cancer Research*, 70(2), 431–435. https://doi.org/10.1158/0008-5472.CAN-09-3049

Slebos, R. J. C., Oh, D. S., Umbach, D. M., & Taylor, J. A. (2002). Mutations in tetranucleotide repeats following DNA damage depend on repeat sequence and carcinogenic agent. *Cancer Research*, 62(21), 6052–6060.

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

Turrina, S., Caratti, S., Ferrian, M., & De Leo, D. (2016). Are rapidly mutating Y-short tandem repeats useful to resolve a lineage? Expanding mutability data on distant male relationships. *Transfusion*, 56(2), 533–538. https://doi.org/10.1111/trf.13368

Tusnady, G. E., Simon, I., Varadi, A., & Aranyi, T. (2005). BiSearch: Primer-design and search tool for PCR on bisulfite-treated genomes. *Nucleic Acids Research*, 33(1), e9. https://doi.org/10.1093/nar/gni012

Vallone, P. M., & Butler, J. M. (2004). AutoDimer: A screening tool for primer-dimer and hairpin structures. *Biotechniques*, 37(2), 226–231. https://doi.org/10.2144/04372ST03

Vollger, M. R., Logsdon, G. A., Audano, P. A., Sulovari, A., Porubsky, D., Peluso, P., & Sanders, A. D. (2019). Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *BioRxiv*, 635037. https://doi.org/10.1101/635037

Vuichard, S., Borer, U., Bottinelli, M., Cossu, C., Malik, N., Meier, V., & Castella, V. (2011). Differential DNA extraction of challenging simulated sexual-assault samples: A Swiss collaborative study. *Investigative Genetics*, 2(1), 11. https://doi.org/10.1186/2041-2223-2-11

Westen, A. A., Kraaijenbrink, T., Clarisse, L., Grol, L. J. W., Willemse, P., Zuniga, S. B., & Weiler, N. E. C. (2015). Analysis of 36 Y-STR marker units including a concordance study among 2085 Dutch males. *Forensic Science International: Genetics*, 14, 174–181. https://doi.org/10.1016/j.fsigen.2014.10.012

Willems, T., Gymrek, M., Poznik, G. D., Tyler-Smith, C., Erlich, Y., & Genomes Project Chromosome, Y. G. (2016). Population-scale sequencing data enable precise estimates of Y-STR mutation rates. *The American Journal of Human Genetics*, 98(5), 919–933. https://doi.org/10.1016/j.ajhg.2016.04.001

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

Young, A. A., Jr. (2011). Comment: Reactions from the perspective of culture and low-income fatherhood. *The ANNALS of the American Academy of Political and Social Science*, 635(1), 117–122. https://doi.org/10.1177/0002716210390316

Zgonjanin, D., Alghafri, R., Almheiri, R., Antov, M., Toljic, D., Vukovic, R., & Petkovic, S. (2017). Mutation rate at 13 rapidly mutating Y-STR loci in the population of Serbia. *Forensic Science International: Genetics Supplement Series*, 6, e377–e379. https://doi.org/10.1016/j.fsigss.2017.09.171

Zgonjanin, D., Alghafri, R., Antov, M., Toljić, D., Almheiri, R., Petković, S., & Vuković, R. (2017). Rapidly mutating Y-STRs population data in the population of Serbia and haplotype probability assessment for forensic purposes. *Forensic Science International: Genetics Supplement Series*, 6, e383–e384. https://doi.org/10.1016/j.fsigss.2017.09.169

Zhang, G. Q., Yang, S. Y., Niu, L. L., & Guo, D. W. (2012). Structure and polymorphism of 16 novel Y-STRs in Chinese Han Population. *Genetics and Molecular Research*, 11(4), 4487–4500. https://doi.org/10.4238/2012.October.11.1

Zhao, J., Bacolla, A., Wang, G., & Vasquez, K. M. (2010). Non-B DNA structure-induced genetic instability and evolution. *Cellular and Molecular Life Sciences, 67*(1), 43–62. https://doi.org/10.1007/s00018-009-0131-2

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.