ARTICLE

A game of hide and seq: Identification of parallel Y-STR evolution in deep-rooting pedigrees

Sofie Claerhout¹ · Michiel Van der Haegen¹ · Lisa Vangeel¹ · Maarten H. D. Larmuseau^{1,2,3} · Ronny Decorte^{1,4}

Received: 6 June 2018 / Revised: 21 November 2018 / Accepted: 27 November 2018 / Published online: 20 December 2018
© European Society of Human Genetics 2018

Abstract

Short tandem repeats on the Y-chromosome (Y-STRs) are common DNA polymorphisms useful for genetic genealogy, population and evolutionary genetics, human genetics, pathology and forensic sciences. It is important to identify all Y-STR variants and to have knowledge of Y-STR mutation rates in order to correctly estimate the time to the most recent common ancestor (tMRCA) between paternally related individuals. When capillary electrophoresis (CE) is performed to analyze genealogical pairs, Y-STR sequence variations remain hidden when the number of repeats is identical. These hidden variations could be due to parallel Y-STR changes or modifications (PM) that occur independently in different lineages leading to alleles with identical number of repeats. In this study, we detect for the first time twelve PM by analyzing 133 males (960 meioses) in extended deep-rooting family pedigrees on 42 Y-STRs. These PM were observed in nine Y-STR loci with mutation rates of at least 5.94×10^{-3} per generation. Sequencing analysis made it possible to distinguish insertions/deletions in different repeat regions revealing the presence of two unique changes in three PM on rapidly mutating and complex Y-STRs DYS724-ab and DYS518. Sequencing unraveled more information concerning the identity of alleles, and increased allelic discrimination possibilities which is of great importance in population genetics and forensic analysis. Limiting the analysis to CE could lead to wrong ancestral allele assumptions, to false negative interpretations and to tMRCA underestimations. These observations highlight the importance and added value of sequencing analysis and suggest a shift in genotyping methods from CE to next generation sequencing.

Introduction

DNA polymorphisms such as short tandem repeats (STRs) occur frequently in the human genome and serve as interesting tools providing numerous applications in interdisciplinary research. STR typing is very useful for

population and evolutionary genetics, human genetics (e.g., stem cell transplantation), pathology and forensic sciences [1–4]. Male specific Y-chromosome STRs (Y-STRs) provide additional applications as it enables the direct separation of male DNA in mixed samples without female DNA interference, useful in for example fetal gender or male lineage determination and forensic sexual assault cases [3, 5]. Furthermore, Y-STRs enable the identification of genealogical patrilineages as the majority of the Y-chromosome lacks recombination driving its inheritance from father to son in a relative conserved manner [4]. These Y-STR patrilineages are particularly useful for population and evolutionary genetics, kinship analysis (e.g., family history and paternity testing) and forensic familial searching. The latter defines a forensic identification method based on Y-STRs to characterize patrilineages in order to identify close or distant male relatives of the unknown perpetrator using the DNA collected at the crime scene [6]. Discrimination of these familial lineages is possible through the presence of Y-STR variants, due to a meiotic change (replication slippage) in DNA sequence in one of the lineages.

Supplementary information The online version of this article (<https://doi.org/10.1038/s41431-018-0312-2>) contains supplementary material, which is available to authorized users.

✉ Sofie Claerhout
sofie.claerhout@kuleuven.be

- ¹ KU Leuven, Forensic Biomedical Sciences, Department of Imaging & Pathology, Leuven, Belgium
- ² KU Leuven, Laboratory of Socioecology and Social Evolution, Department of Biology, Leuven, Belgium
- ³ Familiekunde Vlaanderen vzw., Merksem, Belgium
- ⁴ UZ Leuven, Laboratory of Forensic genetics and Molecular Archaeology, Leuven, Belgium

Y-haplotype comparison from direct father-son couples reveals Y-STR changes, which made it possible to estimate individual Y-STR mutation rates [7]. Another calculation method called the ‘genealogical pair approach’ is based on paternally related male namesakes (men sharing the same surname) in deep-rooted pedigrees separated by a number of generations [8]. This approach enables the deduction of individual Y-STR mutation rates from a large number of meioses with a minimal set of DNA samples [9]. Both autosomal and Y-STR mutation rates have been observed to be correlated with paternal allele transfer, the age of the father and the allele length [7, 9, 10]. High microsatellite variation makes autosomal STR typing less relevant for extended kinship analysis as the difference in number of shared alleles with (un)related individuals fades away over generations [11].

For distant paternal kinship analysis, it is important to identify all Y-STR variants and to have knowledge of exact individual Y-STR mutation rates in order to correctly estimate the time to their most recent common ancestor (tMRCA). To date, various Y-STR markers have been characterized with their mutation rate ranging between 10^{-4} and 10^{-2} per generation (mpg) [7, 9, 12, 13]. The latter group ($>10^{-2}$ mpg) defines the rapidly mutating (RM) Y-STRs and is useful to discriminate close paternally related males [14]. The inclusion of RM Y-STRs in modern Y-STR profiling kits increases the discriminatory power and the weight-of-evidence for a Y-haplotype match [15].

Y-STR typing is still widely performed through fragment analysis by capillary electrophoresis (CE), which characterizes the number of repeats through size separation [6]. CE technology improvements (e.g., novel fluorescent dyes, 5-dyes and 6-dyes) and new PCR-based STR assays increased the throughput and postponed the need for the more laborious and expensive methodology of next generation sequencing [16]. Fragment size allele frequencies are valuable in for example the prediction of useful objectives in forensic medicine and in the observation of the chimeric status after stem cell transplantation [1]. However, when CE genotyping is used to analyze the Y-haplotypes of genealogical pairs, there is a possibility that certain Y-STR variations at sequence level will not be detected, referring to hidden variants. It is for example impossible to distinguish a loss of a tetrameric repeat with a deletion of 4 bp located in the flanking region without sequence analysis [17, 18]. Additionally, sequencing analysis of Y-STRs containing a compound or complex repeat motif (interrupted repeat structures with variable lengths and sequences) can reveal more information concerning the location of the insertion/deletion and therefore increase the observable allelic variation [19, 20].

As Y-STR alleles are rarely sequenced, the question if this lack excludes potential useful information causing false

positive mistakes, remains rather unanswered. An interesting example of hidden STR variations with a possible high impact on kinship analysis are the so called parallel modifications (PM). PM are two (or more) independent DNA sequence slippages during meiotic division resulting into alleles with an identical number of repeats in different lineages of a genealogic tree. The phenomenon of independently originated equal changes has already been observed and described on an evolutionary scale, called homoplasmy or convergent evolution [21]. On an evolutionary scale, it is already known that different Y-SNP haplogroups can have high Y-haplotype resemblance due to recurrent and independent parallel Y-STR changes, causing difficulties in Y-haplogroup age estimations [22] and in studying population genetic patterns [23]. The detection of PM at a genealogical level in deep-rooting family pedigrees could expand our knowledge of homoplasmy to a genealogical time scale. Through sequencing technology, the increased information of allelic variation could eventually reveal a previously hidden PM. If not, at least two unique Y-STR changes remain invisible between two relatives and tMRCA underestimations of biological kinships could be made.

This study arises from previously obtained genealogical pairs collected to investigate the extra pair paternity (EPP) rates and differences in Y-STR mutation rates between Y-haplogroups [8, 9]. Here, we discuss the observation of multiple parallel modification events at a genealogical level in extended family pedigrees including multiple genealogical pairs. Furthermore, we indicate the importance and the added value of detailed sequencing analysis useful for population genetics, genetic genealogy and familial searching.

Materials and methods

DNA samples

This study includes previously obtained genealogical pairs collected to study the EPP rates and haplogroup specific Y-STR mutation rates [8, 9]. Permission for DNA analysis and scientific publication of the anonymized results were received through written informed consents. The Ethical Commission of University Hospital Leuven accepted and approved this study (S55864, S59085). DNA samples were collected, extracted and genotyped as described in Claerhout et al. [9]. All DNA samples were genotyped through CE in four multiplex assays for 42 Y-STR loci, including six RM Y-STR loci (DYS570, DYS449, DYS724-ab, DYS627 and DYS518) to obtain their Y-haplotype, and through SNaPshot® Multiplex System for 81 Y-SNPs to determine their Y-subhaplogroup [9]. More information concerning

the used Y-STR markers and their sequences are available in Supplementary Table 1. A total number of 22 extended deep-rooting pedigrees consisting of at least four patrilineal related males sharing the same surname were included in the present study. These family pedigrees count for 133 males and 960 generations (or meiosis). Y-chromosomal data used in this study has been submitted to the open access Y-STR Haplotype Reference Database (YHRD, <https://yhrd.org>) available under accession numbers YA003651, YA003652, YA003653, YA003739, YA003740, YA003741, YA003742, YA004300 and YA004301. Sequence data obtained in this paper is available in the open access Genbank sequence database (NCBI, <https://www.ncbi.nlm.nih.gov/genbank>) under accession numbers MH814943 to MH814973.

Sequencing of Y-STR alleles

Parallel modifications (PM) were confirmed by DNA sequence analysis of all relatives for the concerning Y-STR loci. DNA samples were amplified by PCR (Applied Biosystems) with identical conditions as the previously mentioned Y-STR analysis [9]. PCR products were purified with ExoSAP-IT™ Cleanup Reagent (Isogen Life Science, The Netherlands) using the standard protocol. DNA sequence analysis was performed using the PCR primers (forward and reverse reaction) and the BigDye® Terminator v3.1 cycle sequencing kit (Applied Biosystems Inc.) according to the manufacturer's specifications. Sequencing products were ethanol precipitated and separated on the ABI PRISM 3130 XL Genetic Analyzer with POP7 and a 50 cm capillary (Applied Biosystems Inc.). DNA sequences were analyzed with Sequencing Analysis v5.2 (Applied Biosystems Inc.) and aligned using the BioEdit Sequence Alignment Editor v7.2.6. All derived sequences were aligned with a reference sequence obtained from the genome browser UCSC (<http://genome.ucsc.edu>) and from a relative within the family pedigree. DNA sequences of all relatives were checked on Y-STR loci with identified PM in order to confirm previous CE genotyping results and to observe sequence composition differences. The sequences of the multi-copy Y-STR DYS724-ab were deduced taking into account the sequence overlap of both copies (Supplementary Figure 1). Additionally, previously collected relatives separated by more than 15 meiosis with an equal number of repeats for Y-STR DYS518 were further investigated at sequence level to unravel more hidden PM.

Origin and implications of parallel modifications

In order to identify the molecular factors influencing the presence of PM, the individual Y-STR mutation rates, the allele size and the average age of the father in all extended

family pedigrees were investigated. Linear regression between the Y-STR mutation rate and the number of identified PM was investigated using R software. The Y-STRs containing PM and their allele sizes were compared with the individual Y-STR mutation rates and average Y-STR allele sizes obtained in our previous study [9]. Statistical significance was tested through the Chi-square test and Fisher's exact test using R software. The average age of the father in the lineage where a PM event took place was compared with the average age of the father of the other generations lacking a PM event. The influence of the age of the father on parallel mutability was explored using a simple two tailed t-test. To frame the consequences of these hidden PM, a case study was simulated to estimate the tMRCA. This analysis was based on comparing different types of information such as CE results, sequencing results and these in combination with additional relatives branching off both paternal lines of the family pedigree. tMRCA estimations and probability curves were obtained through the online tMRCA calculator of J.D. McDonald based on the infinite allele model (IAM) of Walsh (<http://www.scs.illinois.edu/~mcdonald/tmrca.htm>) [24].

Results

Identification of PM

In total, 173 Y-STR changes were observed over 35,520 allele transfers, whereof 24 (13.87%) were observed to be parallel in different lineages. These PM were distributed over five different pedigrees, including 393 meiosis, and were observed in nine different Y-STR loci, namely DYS449, DYS458, DYS518, DYS570, DYS576, DYS627, DYS635 and DYS724-ab (Table 1). These Y-STRs have a mutation rate of at least 5.94×10^{-3} mpg and six of them even higher than 1×10^{-2} mpg referring to RM Y-STRs. Five Y-STR loci have a complex repeat motif consisting of multiple repeats with variable lengths and sequences, and DYS635 has a compound repeat motif containing multiple simple repeats with usually a difference of one nucleotide.

Five PM were identified within the same extended pedigree, further referred to as Pedigree1 (Fig. 1). This pedigree includes 28 judicially related males whereof 24 were identified to be biologically related based on their Y-haplotype and Y-haplogroup. Within Pedigree1, 27 one-step, a two- and a three-step change were observed (Fig. 1). The five PM of Pedigree1 were one-step changes on four different Y-STR loci (DYS724-ab, DYS458, DYS518 (x2) and DYS627). An overview of the other extended pedigrees with identified PM can be found in Supplementary Figure 2.

Sanger sequencing resulted in the confirmation of the number of repeats previously established through CE

Table 1 Y-STRs with identified PM sorted by their mutation rate obtained by Clærhout et al. (2018) including the repeat motif characteristics (type, base pairs and sequence) and average allele sizes [9]

Y-STR	PM	Mutation rate ($\times 10^{-3}$)		Repeat motif			Ref.	Average allele size
		Clærhout et al. (2018)		Type	bp	Sequence		
DYS635	1	5.94		Compound	4	TCTA[4]TGTA[2]TCTA[2]TGTA[2] TCTA[2]TGTA[n]TCTA[n]	(36) ii	22.49
DYS576	1	8.20		Simple	4	AAAG[n]	(36) i	17.42
DYS458	1	9.33		Simple	4	GAAA[n]	(36) i	16.61
DYS570	1	10.18		Simple	4	TTTC[n]	(36) i	17.74
DYS449	1	11.59		Complex	4	TTTC[n]N[50]TTTC[n]	(36) i	29.58
DYS724-ab	3	15.19		Complex	4	CCTT[n]TCCTT[1]CTTT[n]	(37) i,ii	36.29
DYS627	1	15.55		Complex	4	AGAA[3]N[16]AGAG[3]AAAG[n] N[81] AAGG[3]	(6) i	20.84
DYS518	3	18.65		Complex	4	AAAG[3]GAAG[1]AAAG[n]GGAG[1] AAAG[4]N[6]AAAG[n]	(6) i	38.20

Variable sequence motif in bold letter type

i: ISOGG Ybrowse (<http://isogg.org>), ii: complementary sequence

(Supplementary Table 2). Sequencing analysis enabled the discrimination of three identified PM in the complex Y-STR loci *DYS724-ab* (PM1_Pedigree1) and *DYS518* (PM4_Pedigree1 and PM10_Pedigree3) (Table 2). For PM1_Pedigree1, relative 3 and 24 had a parallel loss of one repeat resulting in a 35 repeat length instead of 36. Both changes in PM1 could be identified due to the complex repeat sequence of *DYS724-ab* revealing the different locations of the deletions. Sequencing analysis results for PM1 on the multi-copy Y-STR *DYS724-ab* are shown in Supplementary Figure 1. Another interesting remark for *DYS724-ab* in Pedigree1, is the sequence composition of relative 4. Fragment analysis resulted in 36 repeats as the allele of the MRCA in this pedigree. However, sequencing revealed differences in the number of repeats per repeat structure compared to his close relatives, revealing two additional changes in this lineage on the same locus previously hidden through CE (Table 2). For PM4_Pedigree1 the ancestor allele for *DYS518* was identified to consist of 39 repeats, while relative 13 and 14 changed in parallel to 40 repeats (gain of one repeat). PM4 could be differentiated based on insertions on different locations in their repeat motif. While the repeat insertion of relative 13 was located in the beginning of the sequence, relative 14 had an insertion at the end of the sequence.

For PM10_Pedigree3, it was rather difficult to indicate the exact locations of the PM based on CE results as there were two possible scenarios due to the distribution of relatives (Fig. 2). In the first case, the allele of the MRCA for *DYS518* contains 37 repeats and a parallel gain of one repeat could have occurred in the common patrilineal line of relative 3, 4 and 5 and in the patrilineal line of relative 6 (Fig. 2a). In the second case, the ancestor allele has 38 repeats and a parallel loss of one repeat occurred in the

common patrilineal line of relative 1 and 2 and in the common patrilineal line of relative 7 and 8 (Fig. 2b). However, this dilemma could be resolved through sequence differences: relative 3, 4 and 5 contained the same insertion of one repeat at the end of the sequence, but relative 6 had an insertion of a repeat at the beginning of the sequence (Table 2). Since relative 1, 2, 7 and 8 contained equal sequence compositions, it is certain that the ancestor allele for *DYS518* in Pedigree3 can be identified to consist of 37 repeats. This indicates that PM10 is situated in the common patrilineal lineage of relative 3, 4 and 5 and in the parallel lineage of relative 6, as indicated in Fig. 2a.

Additional results were obtained through sequencing analysis of extra genealogical pairs separated by a large number of generations for the most rapidly mutating and complex Y-STR *DYS518*. In total, 121 previously collected male relatives distributed over 44 deep-rooting pedigrees counting for 1106 meioses were additionally sequenced for *DYS518* (Supplementary Table 3). Despite the fact that the majority revealed no sequence composition differences, it was still possible to reveal one additional PM in a deep-rooting pedigree (PM13_Pedigree28) containing three relatives separated by 28 meioses, see Fig. 3. CE genotyping identified 39 repeats for relative 1 and 38 repeats for relative 2 and 3, revealing a change in the paternal line of relative 1 (Fig. 3a). However, sequencing analysis identified two independent deletion events at different locations of the *DYS518* repeat motif between relative 2 and 3 compared to relative 1 (Fig. 3b). This revealed two unique previously hidden variants, altering the location of the formerly assumed change (Fig. 3c). CE analysis on all 42 Y-STR loci revealed a total amount of five different changes between relative 2 and 3, while in reality at least two extra changes on *DYS518* need to be included.

Fig. 1 Pedigree1 contains 28 judicially related males whereof 24 males are biologically related. Four relatives (crossed) were excluded due to the presence of an interruption in their Y-chromosome lineage. All Y-STR changes are listed whereby the five PM for DYS724-ab,

DYS458, DYS518 (x2) and DYS627 (resp. PM1, PM2, PM3, PM4 and PM5) are indicated in shades of gray; asterisk (*) equal number of repeats as relative 1

Origin of PM

A significant positive correlation between the Y-STR mutation rate and the number of identified PM was observed ($p\text{-value} = 9.5 \times 10^{-11}$). PM allele sizes were observed to be increased by on average one repeat compared to the overall average allele sizes as indicated in Table 1, enlightening a significant positive correlation between PM and allele sizes ($p\text{-value} = 2.7 \times 10^{-6}$). Moreover, when investigating the sequence location of the changes on complex Y-STRs DYS724-ab and DYS518, a significant difference was found between the number of changes observed on long and short repeat motifs ($p\text{-value} = 4.4 \times 10^{-3}$). The longest repeat motif (average = 18.4 repeats) within the sequence underwent 25 changes, while on the shortest repeat motif (average = 14.3 repeats) only 9 changes were present. The average age of the paternal line for the twelve PM are visualized in Supplementary Figure 3. For eight PM, the age

of the father was higher compared to the rest of the pedigree and significant differences of respectively 7.7 years and 6 years were found in PM7_Pedigree2 and PM10_Pedigree3 ($p\text{-values} = 4.5 \times 10^{-2}$ and 1.8×10^{-2}). The overall average age of the father was 1.05 years higher for lineages containing PM compared to the rest of the pedigree.

Discussion

Identification of PM

This paper demonstrates for the first time that homoplasy in Y-STR alleles could also be described as a phenomenon detectable at a genealogical time scale. Through the combination of fragment analysis and deep-rooting extended pedigrees, we were able to identify independently originated, equal STR changes, called PM. PM can be defined as

Table 2 Repeat motif sequence discrimination (underlined) on complex Y-STRs DYS724-ab (PM1_Pedigree1) and DYS518 (PM4_Pedigree1 and PM10_Pedigree3)

	PM	Relative No.	Allele calls	HGVS nomenclature	Y-STR repeat motif sequence
DYS724-ab	Pedigree1	2	36	NC_000024.10:g.26152098	<u>CCTT</u> [17]TCCTT[1] <u>CTTT</u> [19]
		4	36	NC_000024.10:g.26152098	<u>CCTT</u> [18]TCCTT[1] <u>CTTT</u> [18] → 2 modification events in one patrilineage
		3	35	NC_000024.10:g.26152098	<u>CCTT</u> [16]TCCTT[1] <u>CTTT</u> [19]
		24	35	NC_000024.10:g.26152098	<u>CCTT</u> [17]TCCTT[1] <u>CTTT</u> [18]
DYS518	Pedigree1	1	39	NC_000024.10:g.13096074	AAAG[3]GAAAG[1] <u>AAAG</u> [18]GGAG[1]AAAG[4]GAAAGAG[1] <u>AAAG</u> [12]
		13	40	NC_000024.10:g.13096074	AAAG[3]GAAAG[1] <u>AAAG</u> [19]GGAG[1]AAAG[4]GAAAGAG[1] <u>AAAG</u> [12]
		14	40	NC_000024.10:g.13096074	AAAG[3]GAAAG[1] <u>AAAG</u> [18]GGAG[1]AAAG[4]GAAAGAG[1] <u>AAAG</u> [13]
DYS518	Pedigree3	1	37	NC_000024.10:g.13096074	AAAG[3]GAAAG[1] <u>AAAG</u> [15]GGAG[1]AAAG[4]GAAAGAG[1] <u>AAAG</u> [13]
		3	38	NC_000024.10:g.13096074	AAAG[3]GAAAG[1] <u>AAAG</u> [15]GGAG[1]AAAG[4]GAAAGAG[1] <u>AAAG</u> [14]
		4	38	NC_000024.10:g.13096074	AAAG[3]GAAAG[1] <u>AAAG</u> [15]GGAG[1]AAAG[4]GAAAGAG[1] <u>AAAG</u> [14]
		5	38	NC_000024.10:g.13096074	AAAG[3]GAAAG[1] <u>AAAG</u> [15]GGAG[1]AAAG[4]GAAAGAG[1] <u>AAAG</u> [14]
		6	38	NC_000024.10:g.13096074	AAAG[3]GAAAG[1] <u>AAAG</u> [16]GGAG[1]AAAG[4]GAAAGAG[1] <u>AAAG</u> [13]

Variable sequence motif in bold letter type

unique allelic changes in different pedigree lineages both changing to an identical number of repeats. They remain hidden in genealogical pairs when DNA samples are genotyped through PCR-CE, unless at least two other patrilineal relatives are included in the genealogy branching off before the PM events. In our study, 133 genotyped relatives distributed over 22 extended deep-rooting pedigrees revealed twelve PM over nine different Y-STR loci. In 1997, Heyer et al. were pioneers in genotyping a deep-rooted family pedigree of 14 relatives (89 meioses) [25]. Unfortunately, only two unique changes on nine slowly mutating Y-STRs could be observed. Kayser et al. (2007) analyzed two deep-rooted genealogies with 63 meioses and observed five different changes over 68 Y-STR loci [26]. Since our study analyzes 42 Y-STRs with variable mutation rates over a total amount of 960 meioses, it became possible to observe 173 different STR changes, whereof 24 in parallel (13.87%). These PM were detected within nine Y-STR loci with a mutation rate of at least 5.94×10^{-3} mpg, including six RM Y-STR loci ($>10^{-2}$ mpg). The lower number of meioses analyzed in previous studies and the lack of analyzing RM Y-STRs can explain the previously unobserved phenomenon of parallel modification events.

After sequencing all relatives within the family pedigree, the results obtained by CE could be confirmed. For the majority of the PM sequences, it remained impossible to detect any differences at sequence level, which limits further differentiation possibilities. Nevertheless, sequence composition differences were observed within three PM (25%) detected in the complex Y-STRs DYS724-ab (PM1) and DYS518 (PM4 and PM10). As already described in literature, the sequence composition of multiple repeats with variable lengths and sequences in complex STRs provide the opportunity to discriminate allelic changes when they occur on different locations within the STR sequence structure [19]. In PM1, two repeat deletions on different locations of the repeat motif sequence of DYS724-ab made it possible to reveal the previously hidden PM. PM4 and PM10 on DYS518 were differentiated through two repeat insertions on different locations of the repeat motif sequence. The other five PM within complex Y-STRs occurred in the same repeat structure, causing the sequence to be equally aligned. In this case, the extra (or missing) repeat was visualized at the end of the repeated sequence, while in reality it could have occurred anywhere in the repetitive sequence. This type of sequence differences cannot be detected and explains why the discrimination capacity of PM is almost equal to zero in Y-STRs with simple and compound repeats. Three out of twelve PM were identified within the most rapidly mutating and complex Y-STR DYS518 (18.65×10^{-3} mpg), whereof two PM could be

Fig. 2 CE analysis for DYS518 resulted in a dilemma for PM10 in Pedigree3 containing 10 patrilineal relatives whereof two are excluded due to an interruption in their Y-chromosome lineage. **a** The MRCA

allele contains 37 repeats and the location of a parallel gain of repeat is indicated in black. **b** The MRCA allele contains 38 repeats and the location of a parallel loss of repeat is indicated in black

Fig. 3 Assumptions on the position of Y-STR changes in the paternal line based on CE and sequencing of PM13 in Pedigree28. **a** Through CE genotyping a gain of repeat located in the paternal line of relative 1 was assumed. **b** DYS518 repeat motif sequence discrimination (underlined). **c** Sequencing analysis revealed the identification of two different loss of repeats located in the paternal line of both relative 2 and 3

discriminated through sequencing. Additionally investigating the DYS518 sequence of 121 extra male relatives counting for 1106 meioses, enabled the visualization of another PM between two biologically related males separated by 20 meioses (Fig. 3). In total, four PM became distinguishable through sequencing analysis between

genealogical pairs which would remained undetected by CE fragment analysis.

Origin of PM

The occurrence of PM events was significantly correlated with the Y-STR mutation rate. The eight most RM Y-STR loci incorporated in this study contained a PM, which is logical as they are more susceptible to randomly change in different lineages of the pedigree. The question rises if these hidden PM will have an impact on previously deduced individual mutation rates, as PM remain invisible when genotyping only two distantly paternal relatives. A total amount of 21 changes were identified for the Y-STR DYS518 in the dataset of 22 extended family pedigrees, whereof six changed in parallel. A mutation rate of 21.88×10^{-3} (95% CI: 13.59×10^{-3} – 33.24×10^{-3}) was extracted when encountering PM and 15.63×10^{-3} (95% CI: 8.77×10^{-3} – 25.64×10^{-3}) without PM events. As there is no significant difference (p -value = 0.40) between these mutation rates, nor between those previously observed in Claerhout et al. (p -value = 0.52, p -value = 0.53), it can be concluded that formerly established rates do not have to be revised [9].

The average allele size and the age of the father at Y-chromosome transmission were also significantly correlated with the presence of PM. PM allele sizes were observed to be on average one repeat larger compared to the overall average allele sizes. Furthermore, it could be observed that changes occurred more frequently on the longest repeat motif structure within complex Y-STRs. This is reasonable as the mutation rate was already seen to be correlated with

the number of repeats in many studies [7, 27]. The average age of the paternal line for eight out of twelve PM was higher compared to the rest of the pedigree, whereof two were significant higher. These observations are in accordance to previous results concerning molecular factors influencing the replication slippage causing unique sequence changes [9]. Several hereditary neurological disorders are associated with repeat instability (e.g., myotonic dystrophy, fragile X syndrome, Huntington disease, spinal and bulbar muscular atrophy) mostly within repeat motifs containing three to five nucleotides [28–30]. In 2013, Pavicevic et al. described the relation between mutation rate and initial repeat copy number when investigating the molecular genetics of myotonic dystrophy in family pedigrees [31]. When a certain threshold was exceeded, the loci even changed upon every transmission. Additionally, longer repeats were associated with more severe symptoms, an earlier age at onset and the change in expansion size. As the knowledge concerning allelic variation causing several diseases is not yet fully known, it remains necessary to detect all STR changes in order to avoid hidden PM in multiple lineages [32]. For applications as human genetics, genetic genealogy and forensic sciences, it is therefore important to identify all possible allelic variations, and to fully understand the molecular mechanisms of repeat instability.

Implications of PM

The occurrence of PM has consequences in the interpretations of genotyping results which could have an impact on biological kinship analysis used in genetic genealogy and forensic familial searching. Within a genealogical pair, PM remain hidden through fragment analysis, but could be revealed through sequencing when both changes occurred in a different STR repeat structure. Without sequencing (1), the MRCA allele cannot always be assigned leading to MRCA allele dilemmas (2), false negative Y-STR change assumptions could be made causing misinterpretations and (3) all parallel modification events between genealogical pairs remain hidden whereby the tMRCA could be underestimated.

The first problem concerns a dilemma in the identification of the MRCA allele and thus the location in the patrilineage of the parallel modifications. CE genotyping of DYS518 in Pedigree3 point to the presence of a PM (PM10), but the exact patrilineal location of both changes remains unclear due to the distribution of relatives in the pedigree (Fig. 2). After sequencing, two unique insertions became distinguishable as they both occurred in a different repeat structure. This enabled us to identify the patrilineages which changed in parallel for this complex Y-STR and to reveal the length of the MRCA allele.

Second, we encountered that the limitation to CE analysis could eventually lead to false negative interpretations in the patrilineages, as was observed through the discovery of the additional PM (PM13) in Pedigree28 (Fig. 3). Since CE genotyping revealed 39 repeats for relative 1, and 38 repeats for relative 2 and 3, the MRCA allele length would be expected to be 38 and a gain of repeat in the paternal line of relative 1 would be assumed. However, the sequences of the 'equal' 38 repeats revealed two separate deletions at different locations of the DYS518 repeat motif, indicating that the MRCA allele contained 39 repeats. Sequencing thus revealed two different Y-STR variants in the patrilineages where no change was assumed through CE and invalidates the previously assumed Y-STR change. Additionally, sequencing increased the observed number of Y-STR changes between the close relatives 2 and 3 to a total amount of seven. To validate biological kinships between relatives in the pedigree-based dataset, Y-haplotypes were compared through the frequentist approach. Kinship validation was dependent on a match of their Y-SNP sub-haplogroup and on no more than seven differences on 42 Y-STR loci (based on the average mutation rate), as this is stated to be highly improbable for relatives separated by less than 30 generations [24, 33]. Since the incorporation of RM Y-STRs could cross the border of this kinship validation method between genealogical pairs, this limit may thus need to be revised.

The third observed PM consequence were tMRCA underestimations. Two case studies were simulated to estimate the tMRCA between two relatives based on different types of information (Supplementary Figure 4). Case 1 demonstrates tMRCA estimations of two relatives separated by 19 meioses, consequently the tMRCA equals 9 or 10 generations. Through CE genotyping, two differences were observed on 42 Y-STRs estimating their tMRCA to be 3 generations distant. After sequencing, two additional changes were revealed on the complex Y-STR DYS518 (PM10), leading to a correct tMRCA estimation of 9 generations. As fragment analysis missed at least two Y-STR changes between this genealogical pair, it resulted in a false positive close kinship leading to an incorrect tMRCA estimation. Case 2 confirms the added value of sequencing analysis as it visualized two additional unique STR changes (PM10) previously hidden through CE. However, in this case, sequencing analysis could not reveal the other parallel change (PM9) as both events resulted in the same repeat structure. The inclusion of additional relatives to the deep-rooting tree made it possible to observe all six Y-STR changes, revealing the correct tMRCA. Different types of data resulted in diverse numbers of Y-STR changes leading to possible tMRCA underestimations. Despite the fact that sequencing still keeps two STR changes hidden in the last case, it remains

important to include sequence results as the chances of finding two relatives both branching off before the PM in the paternal line are fairly small. Besides, sequence analysis of the complex Y-STR DYS724-ab in Pedigree1 revealed two additional changes in the patrilineage of relative 4, which resulted in an equal amount of repeats as the majority of the pedigree. Taking the other DYS724 locus into account, through sequencing it became possible to observe three independent changes in this RM Y-STR over five generations instead of one. A suitable explanation of this high number could be the high age at Y-chromosome transmission of one of the fathers in his patrilineage (49 years). Again, the lack of sequencing causes different hidden Y-STR changes to remain invisible. These observations are among others important for popular recreational tests used by family historians to find distant relatives and their separated generation time span based on Y-haplotypes [4]. When they rely their results on CE genotyping, the provided tMRCA estimations could possibly be underestimated through the presence of hidden changes. As the online tMRCA calculator of B. Walsh is based on the IAM (every STR change creates a new allele), it does not take PM into account. Despite the fact that he did correct for PM through his stepwise mutation model, it unfortunately does not acknowledge for multi-step STR changes. As the latter has already been described by Ballantyne et al. in father-son pairs and by Claerhout et al. in genealogical pairs, the models could be further improved in order to have more correct tMRCA estimations [7, 9, 24].

To avoid the presence of hidden Y-STR changes, one could exclude genotyping RM Y-STRs as they are more likely to undergo independent changes in different lineages of the pedigree. However, their fast mutation rates provide an increase in Y-STR heterozygosity, giving in theory every (non)paternally related man a different haplotype. As a difference in one locus is sufficient to further distinguish two close relatives, it is very interesting to include multiple RM Y-STRs. In forensic familial searching, the patrilineal inheritance of the Y-chromosome helps to find potential leads to the male offender(s) through the search for relatives. It is therefore important to analyze Y-STRs with slow mutation rates to increase the chances of finding (distant) relatives, and Y-STRs with fast mutation rates to distinguish (close) relatives. As sequencing does not necessarily increase the discrimination capacity of PM within the same sequence structure, the tMRCA could still slightly be underestimated when only two relatives are analyzed. It is therefore important for forensic familial searching to develop a new and improved model with a well-considered choice of Y-STR markers. As the exact tMRCA predictions needed for kinship analysis depend on the detection of the observed Y-STR changes and their mutation rate, it can affect the

timescale from tens to hundreds of generations. In this way, researchers would investigate false close Y-haplotype matches leading to a waste of time and funding as distant related genealogical pairs are very challenging to reconstruct due to the lack of archival information. To enlarge the potential of revealing hidden Y-STR changes between two relatives, it is thus certainly useful to sequence all DNA samples on complex and RM Y-STRs, like DYF387S1-ab, DYS449, DYS724-ab, DYS627, and DYS518.

Conclusion

For the first time, it was possible to detect 24 Y-STR changes in deep-rooting family pedigrees evolving into equal allele lengths, referring to PM. PM remain hidden through the classical CE genotyping method unless multiple relatives were included in the pedigree. Sequencing made it possible to distinguish 44.4% of the PM on complex Y-STR loci as both changes were located in different repeat regions. Therefore sequencing of STR alleles does not only increase the allele discrimination capability useful for population genetics and forensic sciences, but also reveals hidden Y-STR variation in genetic genealogy. These observations further underline the importance of sequencing analysis and urge for a shift in genotyping methods from CE to next generation sequencing.

Acknowledgements We thank all DNA donors of the genetic genealogy project and Familiekunde Vlaanderen vzw for their participation. Funding was provided by KU Leuven (BOF-C1 grant C12/15/013) and the Fund for Scientific Research – Flanders (FWO-Vlaanderen, research grant number 1503216 N).

Author contributions Idea and supervision: SC, RD. Sampling and genealogies: SC, MHD, RD. Genotyping and sequencing: SC, MV, LV. Data analysis: SC. Writing: SC, RD.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Tilanus MGJ. Short tandem repeat markers in diagnostics: What's in a repeat? *Leukemia*. 2006;20:1353–5.
2. Brunstein J. STR typing: method and application. *Med Lab Obs Mlo*. 2014;46:40–2.
3. Barra GB, Santa Rita TH, Chianca CF, et al. Fetal male lineage determination by analysis of Y-chromosome STR haplotype in maternal plasma. *Forensic Sci Int Genet*. 2015;15:105–10.
4. Calafell F, Larmuseau MHD. The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research. *Hum Genet*. 2017;136:559–73.
5. Roewer L. Y chromosome STR typing in crime casework. *Forensic Sci Med Pathol*. 2009;5:77–84.

6. Kayser M. Forensic use of Y-chromosome DNA: a general overview. *Hum Genet.* 2017;136:621–35.
7. Ballantyne KN, Goedbloed M, Fang R, et al. Mutability of Y-chromosomal microsatellites: Rates, characteristics, molecular bases, and forensic implications. *Am J Hum Genet.* 2010;87:341–53.
8. Larmuseau MHD, Claerhout S, Gruyters L, et al. Genetic-genealogy approach reveals low rate of extrapair paternity in historical Dutch populations. *Am J Hum Biol.* 2017;29:1–9.
9. Claerhout S, Vandenbosch M, Nivelte K, et al. Determining Y-STR mutation rates in deep-rooting genealogies: Identification of haplogroup differences. *Forensic Sci Int Genet.* 2018;34 (November 2017):1–10.
10. Sun JX, Helgason A, Masson G, et al. A direct characterization of human mutation based on microsatellites. *Nat Genet.* 2012;44:1161–5.
11. Maguire CN, McCallum LA, Storey C, Whitaker JP. Familial searching: a specialist forensic DNA profiling service utilising the National DNA Database® to identify unknown offenders via their relatives—The UK experience. *Forensic Sci Int Genet.* 2014;8:1–9.
12. Chandler JF. Estimating per-locus mutation rates. *Journal of Genetic Genealogy.* 2006;2:27–33.
13. Willems T, Gymrek M, Poznik GD, Tyler-Smith C, Erlich Y. Population-scale sequencing data enable precise estimates of Y-STR mutation rates. *Am J Hum Genet.* 2016;98:919–33.
14. Ballantyne KN, Ralf A, Aboukhalid R, et al. Toward male individualization with rapidly mutating Y-chromosomal short tandem repeats. *Hum Mutat.* 2014;35:1021–32:e1007028.
15. Andersen MM, Balding DJ. How convincing is a matching Y-chromosome profile? *PLoS Genet.* 2017;13:e1007028.
16. Børsting C, Morling N. Next generation sequencing and its applications in forensic genetics. *Forensic Sci Int Genet.* 2015;18:78–89.
17. Hill CR, Kline MC, Mulero JJ, et al. Concordance study between the AmpFℓSTR® MiniFiler™ PCR amplification kit and conventional STR typing kits. *J Forensic Sci.* 2007;52:870–3.
18. Dauber EM, Kratzer A, Neuhuber F, et al. Germline mutations of STR-alleles include multi-step mutations as defined by sequencing of repeat and flanking regions. *Forensic Sci Int Genet.* 2012;6:381–6.
19. Rockenbauer E, Hansen S, Mikkelsen M, Børsting C, Morling N. Characterization of mutations and sequence variants in the D21S11 locus by next generation sequencing. *Forensic Sci Int Genet.* 2014;8:68–72.
20. Dalsgaard S, Rockenbauer E, Gelardi C, et al. Characterization of mutations and sequence variations in complex STR loci by second generation sequencing. *Forensic Sci Int Genet Suppl Ser.* 2013;4:e218–9.
21. Wake DB, Wake MH, Specht CD. Homoplasmy: from detecting pattern and mechanism of evolution. *Science.* 2011;331 (April):1032–5.
22. de Knijff P. Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. *Am J Hum Genet.* 2000;67:1055–61.
23. Larmuseau MHD, Vanderheyden N, Van Geystelen A, et al. Recent radiation within Y-chromosomal haplogroup R-M269 resulted in high Y-STR haplotype resemblance. *Ann Hum Genet.* 2014;78:92–103.
24. 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;158:897–912.
25. Heyer E, Puymirat J, Dieltjes P, Bakker E, De Knijff P. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum Mol Genet.* 1997;6:799–803.
26. Kayser M, Vermeulen M, Knoblauch H, et al. Relating two deep-rooted pedigrees from Central Germany by high-resolution Y-STR haplotyping. *Forensic Sci Int Genet.* 2007;1:125–8.
27. Brinkmann B, Klitsch M, Neuhuber F, Hühne J, Rolf B. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am J Hum Genet.* 1998;62:1408–15.
28. Gatchel JR, Zoghbi HY. Diseases of unstable repeat expansion: mechanisms and common principles. *Nat Rev Genet.* 2005;6:743–55.
29. Pearson CE, Edamura KN, Cleary JD. Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet.* 2005;6:729–42.
30. Brouwer JR, Willemsen R, Oostra BA. Microsatellite repeat instability and neurological disease. *Bioessays.* 2009;31:71–83.
31. Pavicevic DS, Miladinovic J, Brkusanin M, et al. Molecular genetics and genetic testing in myotonic dystrophy type 1. *Biomed Res Int.* 2013;391821:13.
32. Murillo-Melo NM, Márquez-Quiróz LC, Gómez R, et al. Origin of the myotonic dystrophy type 1 mutation in Mexican population and influence of Amerindian ancestry on CTG repeat allelic distribution. *Neuromuscul Disord.* 2017;27:1106–14.
33. Larmuseau MHD, Vanoverbeke J, Van Geystelen A, et al. Low historical rates of cuckoldry in a Western European human population traced by Y-chromosome and genealogical data. *Proc R Soc B Biol Sci.* 2013;280:20132400.