ARTICLE

Mutability of Y-Chromosomal Microsatellites: Rates, Characteristics, Molecular Bases, and Forensic Implications

Kaye N. Ballantyne,[1] Miriam Goedbloed,[1] Rixun Fang,[2] Onno Schaap,[1] Oscar Lao,[1] Andreas Wollstein,[1,3] Ying Choi,[1] Kate van Duijn,[1] Mark Vermeulen,[1] Silke Brauer,[1,4] Ronny Decorte,[5] Micaela Poetsch,[6] Nicole von Wurmb-Schwark,[7] Peter de Knijff,[8] Damian Labuda,[9] Hélène Vézina,[10] Hans Knoblauch,[11] Rüdiger Lessig,[12] Lutz Roewer,[13] Rafal Ploski,[14] Tadeusz Dobosz,[15] Lotte Henke,[16] Jürgen Henke,[16] Manohar R. Furtado,[2] and Manfred Kayser[1,*]

Nonrecombining Y-chromosomal microsatellites (Y-STRs) are widely used to infer population histories, discover genealogical relationships, and identify males for criminal justice purposes. Although a key requirement for their application is reliable mutability knowledge, empirical data are only available for a small number of Y-STRs thus far. To rectify this, we analyzed a large number of 186 Y-STR markers in nearly 2000 DNA-confirmed father-son pairs, covering an overall number of 352,999 meiotic transfers. Following confirmation by DNA sequence analysis, the retrieved mutation data were modeled via a Bayesian approach, resulting in mutation rates from 3.78×10^{-4} (95% credible interval [CI], $1.38 \times 10^{-5} - 2.02 \times 10^{-3}$) to 7.44×10^{-2} (95% CI, $6.51 \times 10^{-2} - 9.09 \times 10^{-2}$) per marker per generation. With the 924 mutations at 120 Y-STR markers, a nonsignificant excess of repeat losses versus gains (1.16:1), as well as a strong and significant excess of single-repeat versus multirepeat changes (25.23:1), was observed. Although the total repeat number influenced Y-STR locus mutability most strongly, repeat complexity, the length in base pairs of the repeated motif, and the father's age also contributed to Y-STR mutability. To exemplify how to practically utilize this knowledge, we analyzed the 13 most mutable Y-STRs in an independent sample set and empirically proved their suitability for distinguishing close and distantly related males. This finding is expected to revolutionize Y-chromosomal applications in forensic biology, from previous male lineage differentiation toward future male individual identification.

Introduction

The nonrecombining part of the human Y chromosome (NRY) is widely used in human population[1] and forensic genetics[2] because it shows a male inheritance and substantial structuring in human populations.[3] With its particular susceptibility to genetic drift caused by low effective population size[4] and the additional influence of patrilineal cultural practices,[5–7] the NRY provides the strongest genetic differentiation over geographic distance when compared with other parts of the genome.[8,9] This has made the NRY exceptionally valuable for the reconstruction of human population history,[9,10] including estimation of demographic parameters,[11] as well as for genealogical relationships[12] and male lineage determination in forensic applications.[13–15] However, all inferences from NRY data need to apply a specific set of models for both the mutation process and the mutation rate assumed to underlie the particular NRY markers used.[16] Commonly, Y-chromosomal microsatellite or short tandem repeat (Y-STR) variation is used to infer temporal and spatial origins of the Y chromosome, particularly the nodes of a phylogenetic tree constructed from single-nucleotide polymorphism (SNP) or DNA sequences.[17,18] As such, evolutionary inferences on timescales of tens to hundreds of generations, as usually applied, are highly dependent on the accuracy of the Y-STR mutation rate estimates used. Furthermore, for forensic applications of Y-STRs such as paternity testing, including deficiency cases involving male offspring and deceased alleged fathers,[19] accurate knowledge of the mutability of the applied Y-STRs is needed to obtain reliable paternity probabilities. Such knowledge is also essential in genealogical studies aiming to establish the relationship between putatively closely or distantly related males.[12]

However, current information about Y-STR mutability is limited, because empirical data are only available for a small set of particular loci. Commonly, either small

[1]Department of Forensic Molecular Biology, Erasmus University Medical Center Rotterdam, Rotterdam 3000 CA, The Netherlands; [2]Life Technologies, Foster City, CA 94404, USA; [3]Cologne Center for Genomics, University of Cologne, Cologne D-50674, Germany; [4]Netherlands Forensic Institute, Den Haag 2497 GB, The Netherlands; [5]Department of Forensic Medicine and Center for Human Genetics, University Hospitals K.U. Leuven, Leuven 3000, Belgium; [6]Institute of Legal Medicine, University Hospital Essen, Essen D-45122, Germany; [7]Institute of Legal Medicine, University of Schleswig-Holstein, Kiel D-24105, Germany; [8]Forensic Laboratory for DNA Research, Department of Human and Clinical Genetics, Leiden University Medical Center, Leiden 2300 RC, The Netherlands; [9]Centre de Recherche, CHU Sainte-Justine, Département de Pédiatrie, Université de Montréal, Montreal PQ H3T 1C5, Canada; [10]Interdisciplinary Research Group in Demography and Genetic Epidemiology, Université du Québec à Chicoutimi, Chicoutimi PQ G7H 2B1, Canada; [11]Abteilung Myologie, Experimental and Clinical Research Center, Charité-Universitätsmedizin Berlin, Berlin 10117, Germany; [12]Institute of Legal Medicine, University of Leipzig, Leipzig 04103, Germany; [13]Abteilung für Forensische Genetik, Institut für Rechtsmedizin und Forensische Wissenschaften, Charité-Universitätsmedizin Berlin, Berlin 10115, Germany; [14]Department of Medical Genetics and Department of Forensic Medicine, Medical University Warsaw, Warsaw 02-007, Poland; [15]Department of Forensic Medicine, Wrocław Medical University, Wrocław 50-368, Poland; [16]Institut für Blutgruppenforschung LGC GmbH, Köln 50933, Germany
*Correspondence: m.kayser@erasmusmc.nl
DOI 10.1016/j.ajhg.2010.08.006.

pedigrees (both deep-rooting and immediate families) or observed repeat variation between isolated human populations has been used to estimate Y-STR mutation rates.[20–22] However, population diversity-based estimates are often indirectly assumed with the help of calibration dates from other sources, such as archeological investigations.[23] Usually, and for the limited set of Y-STRs studied so far, resulting rates are a magnitude lower than family-based rates, which is explained by noninclusion of multistep mutations and back mutations, as well as variation in calibration dates.[23] The more accurate method of estimating Y-STR mutation rates is the direct observation of transmission between father and son, as long as large numbers of genetic transfers (meioses) are covered by testing a large number of father-son pairs. However, reasonably large family data are only available for a small number of particular Y-STRs often used for forensic purposes.[24–28] A recent large study on 17 Y-STRs, which also provided a summary of the most relevant published data covering over 135,000 meiotic transfers, revealed variation in the mutation rates between loci of about 1 magnitude from 2×10^{-4} (95% credible interval [CI], 2×10^{-5} to 8×10^{-3}) to 6.5×10^{-3} (2.3×10^{-3} to 1.3×10^{-2}) per locus per generation.[28] Such noticeable variation in mutation rates between just 17 loci predicts that even higher variation in mutation rates will be found when increased numbers of Y-STRs are examined. However, the lack of reliable mutation rate data for most of the currently known Y-STRs[29] precludes their accurate use for evolutionary inference of population parameters, as well as for others, such as forensic applications.

Just as there is a lack of accurate mutation rate data, there is a lack of consensus regarding the molecular causes of Y-STR mutations because of the limited number of loci studied thus far. Although most research on autosomal STRs confirms that the stepwise mutation model (SMM)[30] is too simplistic to explain the lack of long STRs, questions remain about the exact mechanism in operation for STRs in general. Mutation biases between alleles are commonly observed, with increasing repeat numbers increasing the probability of mutation.[31–33] A proportional bias of expansion versus contraction mutations appears to operate, with longer alleles tending to contract and vice versa,[32–34] although the strength of this relationship is uncertain. It has also been postulated that point mutations play a large role in maintaining a stationary distribution of allele lengths, preventing the infinite growth of repeat segments.[35–38] More recently, the sequence motif itself has been suggested as a major contributing factor to the differences in mutation rates between loci, with specific motifs appearing to mutate more rapidly,[38] with higher heterozygosities between human populations[39] and greater sequence diversity between humans and chimpanzee STRs.[40] However, most conclusions regarding the causes of STR mutation have been formed from either comparative genomic analyses[35,37,40] or indirect polymorphism analyses,[31,39] both of which may miss substantial numbers of mutations. Instead, sequence-based analysis of a large number of Y-STR mutations would allow a more direct investigation of the molecular processes in action. The strict paternal inheritance of STRs on the NRY allows the unequivocal determination of the mutational event in father-son pair studies, which is difficult for autosomal STRs in family studies. Thus, using Y-STRs allows the retrieval of more accurate knowledge about STR mutability in general.

Furthermore, Y-STR markers currently applied to evolutionary, genealogical, and forensic studies have low to midrange mutation rates,[27,28] which makes them ideal tools to distinguish male lineages (i.e., groups of closely and distantly related males sharing almost identical Y chromosomes) in applications involving comparatively recent timescales.[12,15] However, these particular Y-STR markers usually fail to differentiate members of the same male lineage, and as such, the current forensic use of NRY suffers from the strong limitation that conclusions cannot be made on an individual level, as is usually required in forensic investigations. Also, for microevolutionary studies, investigating male genealogies for historical and other purposes,[12] or for investigating histories of populations that underwent strong bottleneck or founder effects,[9,41] the amount of diversity offered by currently used Y-STRs with midrange mutation rates is usually not sufficient. One could speculate that if Y-STRs with substantially higher mutation rates than are currently known for the limited number of markers investigated were available, it may become possible to differentiate male relatives at the individual level, which would solve the current dilemma of Y chromosome applications in forensics.

To address three main issues—(1) the lack of knowledge on Y-STR mutability based on a reasonably large number of loci, as required for evolutionary and genealogical applications, (2) the limited knowledge about the molecular basis of Y-STR mutability, and (3) the lack of Y-STRs for familial differentiation in forensic, genealogical, and particular population applications—we have investigated 186 Y-STRs in ~2000 DNA-confirmed father-son pairs. We not only describe in this study mutation rates and characteristics for the largest number of different Y-STRs ever studied so far, including the first mutation rate estimates for most of these markers, but we also use the diversity and DNA sequence data generated for all loci to investigate the underlying causes of Y-STR mutability. Finally, we empirically tested the suitability of the identified most mutable Y-STRs for male relative differentiation, as well as their implication for Y chromosome applications in forensic science.

Material and Methods

DNA Samples

All father-son pairs used in the mutation rate study were confirmed in their paternity by molecular analyses, utilizing autosomal STRs, Y-STRs, HLA and RFLP genotyping, and blood

grouping, in addition to familial or governmental documentation. A threshold for paternity probability of 99.9% was set for inclusion in the study. Samples were obtained from the Berlin, Leipzig, and Cologne areas of Germany and the Warsaw and Wroclaw areas of Poland. Whole-genome amplification (WGA) with the GenomiPhi DNA Amplification kit (GE Healthcare) was performed on the Leipzig samples because of low DNA quantities. WGA reactions were performed as recommended by the manufacturer, and products were purified with Invisorb 96 Filter Microplates (Invitek GmbH). To verify the value of the smaller set of RM Y-STRs, we obtained an additional independent set of samples from male relatives from the Greifswald, Kiel, and Berlin areas of Germany, the Leuven area of Belgium, the Warsaw area of Poland, and Canada and Central Germany, as described elsewhere.[12] All families and pedigrees were confirmed by the same methods as the father-son pairs; pairs with complete genotypes for both the rapidly mutating (RM) Y-STRs and Yfiler Y-STRs were considered for analysis, or, in the case of partial genotypes, only those that showed a mutation at one or more loci were included. The use of all samples for the purpose of this study was in agreement with the institutional regulations and was under informed consent.

Y-STR Markers and Genotyping Protocols

Y-STR markers were mostly selected from a previous study detailing a large number of 167 previously unknown Y-STRs,[29] with the additional inclusion of Y-STRs known at the time of project commencement.[42] The focus was on single-copy Y-STR markers in order to be able to fully confirm genotype differences by DNA sequence analysis when identifying mutations. However, given our aim to find RM Y-STRs, we included some additional multicopy Y-STRs, especially those with high diversities (for which mutation confirmation was performed by independent genotyping). A complete list of loci, primer sequences, and protocols can be found in Table S1 available online. Seventeen of the 186 Y-STRs were genotyped with a commercially available kit, the AmpF*l*STR Yfiler PCR Amplification kit (Applied Biosystems), following the manufacturer's instructions. Full descriptions of protocols and markers can be found in [28]. The remaining 169 Y-STRs were genotyped via 54 multiplex assays, including 1–5 markers each. PCRs were performed via three differing protocols, and details are provided in Table S1. In addition, 13 Y-STRs identified during the study as RM Y-STRs were genotyped via three multiplex assays in an independent sample set of male relatives. All PCRs were performed on GeneAmp PCR System 9700 machines (Applied Biosystems) at the Department of Forensic Molecular Biology, Erasmus MC Rotterdam. Fragment length analysis was performed with the 3130xl Genetic Analyzer (Applied Biosystems). Profiles generated were genotyped with GeneMapper software (ID v. 3.2, Applied Biosystems). Genotype differences were identified with in-house-developed Microsoft Excel 2007 macros. All mutations were confirmed by DNA sequence analysis in Rotterdam of both the father and son at the Y-STR locus, as described in [28]. Multicopy Y-STR loci with three or more alleles were not able to be sequenced, but mutations were confirmed by at least two independent fragment length analysis amplifications.

Statistical Data Analyses

Mutation rates for individual markers were estimated via a binomial hierarchical Bayesian model[43] with the Marcov chain Monte Carlo (MCMC) Gibbs sampling, as implemented in WinBUGS[44] and as described in [28]. In brief, we assumed that each mutation rate could be considered as a realization of the mutation rate underlying any Y-STR. We assumed that the mutation rate θ_i of Y-STR i was a sample from a common population distribution defined by hyperparameters ϕ. In that way, the estimated mutation rate of a Y-STR incorporates the information provided by the observed data on that Y-STR (number of observed mutations over all the observed father-son pairs) and the information of the mutation rate of "the Y-STR," as estimated in the hyperparameter from all the Y-STRs. In practice, this implies that all Y-STRs will show a mutation rate greater than zero when estimated from the posterior distribution, but the rate will be smaller for Y-STRs where few or no mutations were observed compared to Y-STRs where large numbers of mutations were seen. The mutation rate of each Y-STR was coded in a logit form and assumed to follow a normal distribution with parameters μ and $\tau = 1/\sigma$ to be estimated, as well as the particular mutation rates of each STR. Because only very limited data were available prior to our study for the range of Y-STR mutation rates, we assumed diffuse, noninformative prior distributions for the hyperparameters. A noninformative prior normal distribution ($\mu = 0, \tau = 1 \times 10^{-6}$) was specified for the hyperparameter μ, and a prior diffuse gamma distribution with parameters $\alpha = 1 \times 10^{-5}$ and $\beta = 1 \times 10^{-5}$ was specified for the parameter τ. Three MCMC chains that used the Gibbs sampler were generated in parallel when estimating the mutation rate for each locus, with 100,000 runs performed for each chain. Mean, median, and 95% CI were estimated from the three chains after discarding the first 50,000 runs and performing a thinning of 15 in order to reduce the amount of autocorrelation between adjacent simulations. Locus-specific differences in mutation rates between the sampling populations (Cologne, Berlin, Leipzig, Warsaw, and Wroclaw) were tested by means of a permutation analysis. The average mutation rate for each locus and each population was compared to a hypothetical permutated population in which each father-son pair had been assigned to a population at random, maintaining the original sample sizes for each locus. The number of times the permutated averaged mutation rate was larger than the observed rate was recorded and used to obtain the one-tail p value over 100,000 iterations. The lack of significant differences between populations allowed pooling of mutation rates across populations.

In order to investigate the mutation rate of the Yfiler and RM Y-STR sets rather than of each marker within the set, we computed the total number of mutations observed between each father-son pair for each set, given the number of Y-STRs analyzed. We then modeled this parameter under the Bayesian paradigm with a Poisson distribution. A prior with a Gamma distribution was used[43] with a diffuse shape of 1 and a scale of 200, implying a mutation rate with a mean of 0.005 and a variance of 40,000. The posterior distribution followed a conjugate Gamma distribution with shape of 1 + (total number of mutations) and scale of 1/(1/(200 + total number of markers used)). In order to estimate the probability of observing at least one mutation in each set, 100,000 Monte Carlo replicates were performed with the rgamma function of the R package (see Web Resources) from the estimated shape and scale of the posterior distribution of each set of Y-STRs.

The probability of observing at least one mutation (k) within either of the Y-STR sets in any given father-son pair was directly estimated from the Poisson distribution $P(k > 0) = 1 - P(k = 0) = 1 - e^{-Nm}$, with N representing the number of markers and m representing the average mutation rate of the set of markers obtained from the sampling from the posterior distribution.

The molecular factors determining mutation rates were modeled via a Poisson regression with in-house-developed Matlab scripts (v. 7.6.0.324, The Mathworks). The mutation rate was modeled as a function dependent on the repeat length, the sequence motif, the complexity of the locus,[29] and the length of the repeat in base pairs (tri-, tetra-, penta-, or hexanucleotide), as

$$p(y \mid \theta) = \prod_{i=1}^{n} \frac{1}{y_i!}(x_i\theta)^{y_i} e^{-xi\theta},$$

where θ is assumed to be dependent on the factors described above, in the form $\theta = e^{\alpha L + \beta S + \gamma C + \delta V + \varepsilon R}$, in which L represents the length of the allele (number of repeats, either of the longest homogenous array or the total locus), S represents the sequence motif (comprised of the number of A, T, C, or G nucleotides in the repeated sequence motif), C represents the complexity of the locus, either in binary or quantitative form, V is the number of variable motifs present, and R is the repeat length A stepwise regression procedure was used, with probability to enter ≤ 0.05, probability to remove ≥ 0.10.

For clarity, the methods used for defining and calculating the number of repeats within a locus, as well as the complexity of that locus, are elucidated below.

Locus designations were modeled after [29], in which at least three consecutive repeats of the same motif are required to define a given repeat segment as a locus, and any interruption of more than 1 base, but less than a full unit, is classed as ending the locus. Individual Y-STR loci contained between one and five repeat blocks, for example, DYS612 with five blocks $(CCT)_5(CTT)_1(TCT)_4(CCT)_1(TCT)_{19}$. If a locus contained more than one variable segment and repeat numbers could not be assigned to all individuals at all repeat segments accurately, the locus was removed from the regression analysis. A segment was defined as variable if a variation in repeat number was seen in any individual sequenced, relative to the remainder of the population.

Number of Repeats

The number of repeats in the longest homogenous array was directly counted, and the population average was calculated for each locus. In addition, any additional repeats around the longest array were added to calculate the total number of repeats for each locus. In the above example for DYS612, the length of the longest array is 19 and the total number of repeats is 30.

Repeat Length

The length in base pairs of the repetitive motif ranged from 3 to 6 (including tri-, tetra-, penta-, and hexanucleotide repeats).

Complexity

Two complexity statistics were calculated per locus. First, a binary classification system was used, in which loci with only one repetitive segment (e.g., $(GATA)_{10}$) were classified as simple, whereas any locus with two or more repetitive segments consisting of more than three consecutive repeats (e.g., $(GATA)_{10}(CATA)_3$) was complex. Second, more quantitative information was provided by the complexity formula in Kayser et al.[29]:

$$C = \frac{n^2}{(n-1)^2}\left(1 - \sum_{i=1}^{m}\left(\frac{s_i}{n}\right)^2\right)\left(1 - \sum_{i=1}^{l}\left(\frac{b_i}{n}\right)^2\right),$$

where n is the total number of repeats in the locus, s_i is the number of repeats of the ith sequence motif, and b_i is the number of repeats in the ith block.

Correlation and log-linear regression analyses were carried out in SPSS v. 15.0, as were all mean comparison tests (utilizing analysis of variance, Mann-Whitney U test, and Kruskal Wallis test).

Results

Mutation Rates of Y-STR Markers

We screened 186 tri-, tetra-, penta-, and hexanucleotide Y-STR markers[29,42] for mutations in up to 1966 DNA-confirmed father-son pairs per marker by multiplex fluorescence-based fragment length analysis, giving direct observation of 352,999 meiotic transfers (for technical details, see Table S1). To confirm mutations, we confirmed all Y-STR genotype differences observed between fathers and their sons by DNA sequence analysis for single-copy and duplicated markers or by duplicate fragment length genotyping analysis for multicopy Y-STRs with more than two copies (where sequence analysis was not informative). Overall, we identified 924 confirmed mutations at 120 (64.5%) of the 186 Y-STR markers studied (details of each mutation observed can be found in Table S2). For 66 Y-STR markers, the up to 1966 father-son pairs analyzed did not allow us to detect mutations as a result of a very low underlying mutation rate. The large number of Y-STR markers employed identified the range of Bayesian-based mutation rates estimated from the median of the posterior distribution to be between 3.81×10^{-4} (95% CI, 1.38×10^{-5} to 2.02×10^{-3}) and 7.73×10^{-2} (6.51×10^{-2} to 9.09×10^{-2}) per marker per generation (Figure 1; Table S1). Ninety-one Y-STR markers (48.9%) had mutation rates in the order of 10^{-3}, a further 82 markers (44%) in the order of 10^{-4}, and 13 (6.9%) in the order of 10^{-2}. Across all 186 Y-STR markers, the average mutation rate was 3.35×10^{-3} (95% CI, 1.79×10^{-3} to 6.38×10^{-3}), with an average rate of 4.26×10^{-3} (95% CI, 2.38×10^{-3} to 7.60×10^{-3}) for the 122 tetranucleotide repeats as the largest repeat-length subgroup of Y-STR markers included here. Notably, the 13 Y-STR markers with mutation rates above 1×10^{-2}, representing only 7% of the markers studied (termed RM Y-STRs), covered a large number of 462 of the 924 (50%) mutations observed in the study.

Y-STR Mutation Characteristics

The large number of mutations and the DNA sequence data generated for their confirmation allowed an in-depth examination of the mutation characteristics. However, six Y-STR markers had to be excluded from this type of data analysis because of the presence of multiple variable repeats within the amplicon that prevented unambiguous assignment of repeat length in nonsequenced individuals, giving 787 mutations at 181 Y-STRs. A slight excess of 423 repeat losses over 364 repeat gains was observed, resulting in a repeat loss:gain mutation ratio of 1.16:1 (95% binomial CI, 1.08:1–1.24:1), although the difference was not statistically significant ($t = -1.543$, $p = 0.125$). The vast majority of 757 mutations were single-repeat changes, with only 30 multirepeat changes observed, giving a statistically significant single:multirepeat mutation ratio of 25.23:1 (95% CI, 37.62–17.52, $Z = -9.33$, $p = 1.1 \times 10^{-20}$). Of the 30 multirepeat mutations, 25 were double-step mutations (2 repeat units), 3 were triple-step mutations,

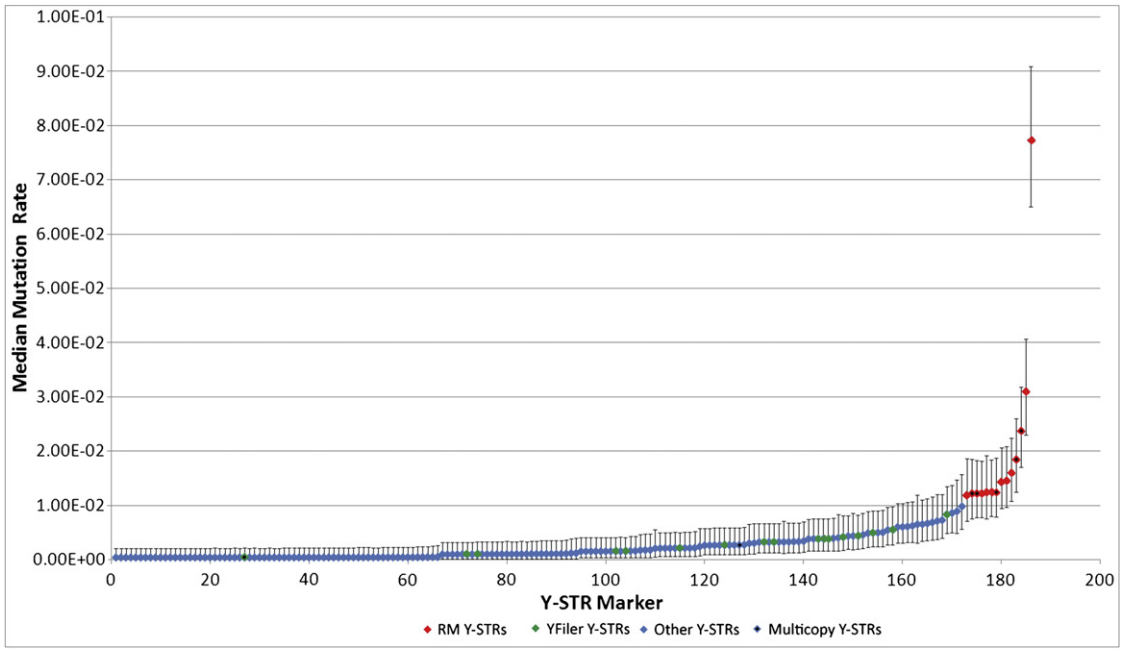

Figure 1. Mutation Rates of 186 Y-STR Markers Established from Father-Son Pair Analysis
Distribution of 186 Y-STR markers according to their Bayesian-based mutation rates (with credible intervals) estimated from analyzing up to 1966 DNA-confirmed father-son pairs per marker. The 13 RM Y-STR markers ascertained for further family or pedigree analysis are highlighted in red, and the commonly used 17 Yfiler Y-STRs are in green. Multicopy Y-STRs are noted with a black insert diamond.

1 was a quadruple-step mutation, and 1 was a quintuple-step mutation (5 repeat units) . Among the multistep mutations, a substantial and statistically significant excess of losses was observed, with 24 multistep losses to 6 multistep gains ($\chi^2 = 29.0$, $p = 7.2 \times 10^{-8}$). Apparent locus duplications between single father-son pairs were found at DYS462 and DYS611, two Y-STR markers normally observed in single copy.

Molecular Factors Influencing Y-STR Mutability

Large mutation numbers and the associated DNA sequence data also allowed us to investigate molecular factors influencing Y-STR mutability. In contrast to the mutation rates quoted above for the Y-STR markers, i.e., PCR amplicons that may include more than one Y-STR locus, we corrected in the following analyses mutation rates for the number of Y-STR copies and loci present. For example, although DYF387S1 is given a single Y-STR marker mutation rate, it has two copies, hence it consists of two Y-chromosomal loci. Because we cannot know, based on our genotyping protocol, at which locus a mutational event within such multicopy marker system occurred, we averaged the mutation rate for each copy and treated them as separate loci. Furthermore, we also separated Y-STR loci within Y-STR markers, in accordance with the rules defined by Kayser et al.[29] in which repetitive sequences that were separated by a nonrepetitive sequence of >1 bp insertion, deletion, or substitution were designated as separate loci. This was done assuming that the mutational process is independent under such criterion. This resulted in a set of 267 Y-STR loci, with 787 mutations observed across 448,824 allelic transfers used in the subsequent analyses.

The effect of specific molecular features on the mutation rates of Y-STR loci was tested by means of a Poisson regression model, including (1) the average number of repeats in the longest homogenous array, (2) the effect of any additional nonvariable repeats directly surrounding the longest homogeneous array, (3) the complexity of the locus (either as a binary simple versus complex model or by using the complexity statistic described above), (4) the length in base pairs of the repetitive motif, and (5) the sequence of the repetitive motif. In cases in which two factors encompassed the same source of variation (with one factor including additional information), separate models were compared for each; for example, the repeat number of the variable repetitive array and total repeat number for the locus were not combined in the same model but were instead compared separately with the mutation rate. The majority of information regarding Y-STR locus mutation rates was contained in a model that included the total number of repeats (including both the longest homogenous repetitive array and any adjacent nonvariable repeat strings of ≥ 3 units), the length of the repetitive motif, and the quantitative complexity of the locus ($\chi^2_{87} = 3.54 \times 10^6$, $p < 0.0001$). This combined model accounted for 87.9% of the variation in mutation rates observed (partial $\eta^2 = 0.879$, $F = 12.86$, $p < 0.0001$); we will describe these three features in more detail below.

Number of Repeats

Two estimates of the average number of repeats were calculated for each Y-STR locus: (1) the average repeat number in

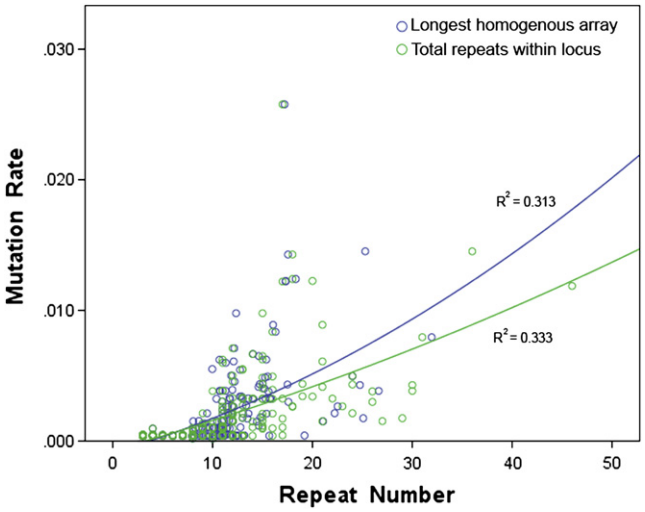

Figure 2. Correlation between the Length of the Longest Homogeneous Array, or the Total Number of Repeats within a Locus, and the Allele-Specific Mutation Rate from 267 Y-STR Loci

Although the number of repeats present within a locus's longest homogenous array can be used to predict mutability, the total number of all repeats present within the locus has higher predictive value.

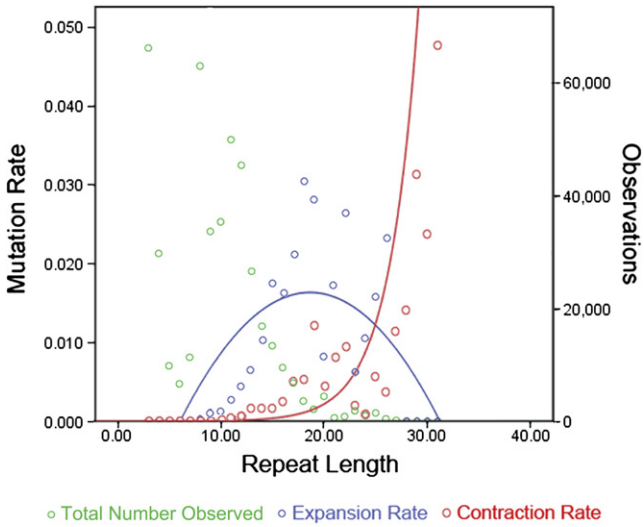

Figure 3. Relationship between Total Number of Repeats and Mutation Direction and Rate from 267 Y-STR Loci

Repeat loss mutations (contractions) displayed an exponential relationship with the total number of repeats, with increasing loss rates at loci with higher numbers of repeats. Repeat gain mutations (expansions) showed a weak quadratic function, with a peak in gain rate at 19 total repeats.

the longest homogenous array, and (2) the repeat number of the longest homogeneous array plus any nonvariable repeats immediately adjacent (in accordance with previously defined rules for motif structure[29]). Our regression analysis showed that, although the number of repeats in the longest homogenous array did influence the mutation rate significantly, with higher numbers of repeats increasing the mutation rate (Wald $\chi^2 = 2.41 \times 10^6$, $p < 0.0001$), including the number of nonvariable repeats surrounding the array provided slightly more accurate information to the model (Wald $\chi^2 = 3.03 \times 10^6$, $p < 0.0001$; Figure 2). The effect size within the model was estimated with a partial η^2 of 0.798, indicating that the variance in the total number of repeats between loci accounts for ~78% of the overall (effect + error) variation in Y-STR mutation rates observed. In addition, a statistically significant exponential relationship was observed between the total number of repeats and the allele-specific mutation rate ($R^2 = 0.707$, $p = 6.84 \times 10^{-9}$). In addition, there was a strong relationship between the total number of repeats and the direction of mutation (Figure 3). Longer alleles displayed an exponential and statistically significant tendency toward repeat losses (contractions; $R^2 = 0.585$, $p = 8.27 \times 10^{-7}$), whereas shorter alleles gained repeats (expansion) significantly more frequently ($R^2 = 0.238$, $p = 0.011$). The expansion mutation rate had a quadratic distribution, with a vertex around 19 repeats.

Complexity of Repetitive Structure

Within the data set of 267 Y-STR loci examined, 193 were simple Y-STR loci, i.e., consisting of only one uninterrupted, homogeneous repeat stretch, and 74 were complex Y-STR loci, i.e., consisting of more than one repeat stretch or a homogeneous repeat stretch interrupted by 1 bp. A statistically significant difference in mutation rates was seen between the simple and complex Y-STR loci ($\chi^2 = 12.377$, $p = 0.0004$), with complex Y-STR loci expressing a higher average mutation rate (2.40×10^{-3}, 95% CI, $1.07 \times 10^{-3} - 5.15 \times 10^{-3}$) than simple ones ($1.65 \times 10^{-3}$, 95% CI, $7.03 \times 10^{-4} - 3.98 \times 10^{-3}$). Furthermore, the ratio between simple and complex Y-STR loci markedly changed depending on the mutation rate observed. Among Y-STR loci with mutation rates in the order of 10^{-4}, the simple: complex Y-STR locus ratio was 1:0.27; loci with rates of 10^{-3} had a 1:0.63 ratio, and loci with rates of 10^{-2} had a 1:0.75 ratio. From this skewed ratio, it could be concluded that the complexity of the locus influenced the mutation rate, and as such, measures were included in the regression model. Furthermore, we applied two statistics to measure the effect of Y-STR locus complexity on the Y-STR mutation rate. Initially, a binary simple versus complex classification was used, and although this provided statistically significant information to the model (Wald $\chi^2 = 5.19 \times 10^5$, $p < 0.0001$), it was not as informative as the subsequently applied quantitative complexity estimate derived by Kayser et al.[29] In this, all repeats within a locus were used to estimate the sequence complexity quantitatively. Loci with more blocks of different repeat motifs have higher complexities, as do loci with equal numbers of repeats between the different motifs. The total amplicon complexity provided significantly more information to the regression model than the binary classification (Wald $\chi^2 = 9.22 \times 10^5$, $p < 0.0001$). The partial η^2 estimate of 0.644 suggests that the complexity of a locus contributes substantially to the variation in mutation rates between loci.

Length of Repetitive Motif

The third parameter influencing Y-STR mutability was the length in base pairs of the repetitive motif, with a statistically significant decrease in mutation rate as the repeat length increased (Wald $\chi^2 = 6.111 \times 10^4$, $p < 0.0001$). However, the low number of penta- and hexanucleotide repeats on the one hand and the absence of dinucleotide repeats on the other may give a bias in this analysis, as observed by the low partial η^2 estimate of effect size at 0.012. The average mutation rate was 1.11×10^{-3} (95% CI, 6.24×10^{-4} to 1.59×10^{-3}) for the 72 trinucleotide Y-STR loci, 2.07×10^{-3} (95% CI, 1.60×10^{-3} to 2.54×10^{-3}) for the 175 tetranucleotide Y-STRs, 1.55×10^{-3} (95% CI, 1.03×10^{-3} to 2.07×10^{-3}) for the 17 pentanucleotide Y-STRs, and 4.64×10^{-4} (95% CI, 4.20×10^{-4} to 4.71×10^{-4}) for the 3 hexanucleotide Y-STRs. Although small, the differences in average mutation rates between the repeat length categories proved to be statistically significant ($Z = -14.165$, $p < 0.0001$).

Father's Age

We also tested, as an additional factor exclusive of the regression model, the influence of the father's age at the time of the son's birth to Y-STR mutability. The average father's age without Y-STR mutations observed was 30.55 ($\pm$10.73) years, compared to 32.42 ($\pm$10.97) years for fathers with at least one observed mutation, and the difference was highly statistically significant ($Z = -5.27$, $p = 1.37 \times 10^{-7}$). We also observed a small but statistically significant positive correlation between the number of mutations observed at Y-STR markers and the age of the father ($R^2 = 0.141$, $p = 1 \times 10^{-6}$).

Nonsignificant Predictors of Y-STR Mutability

Intriguingly, the number of variable motifs and the sequence of the repetitive motif did not turn out to contribute sufficient information to the Y-STR mutability model. For the number of variable motifs present within a given Y-STR locus, this may have been influenced by the way a locus was defined—very small stretches of interrupting sequence would split a sequence into two separate loci. As such, there were very few (3.7%) loci that had more than one variable array, resulting in a small effect size (partial $\eta^2 = 0.0001$) and nonsignificance in the model (Wald $\chi^2 = 0.180$, $p = 0.671$). The influence of the sequence motif of the main repeat (composing the longest homogenous array) toward Y-STR mutability was also tested. Following Kelkar et al.,[40] the repetitive motif was classified into 8 sequence types for trinucleotide repeats, 16 for tetranucleotide repeats, and 24 for pentanucleotide repeats. These repeat designations were based on the number of A nucleotides in the repetitive motif, regardless of the strand direction; for example, CTTC, TCTC, and GAAG were all classified as AAGG repeats. However, the sequence motif did not prove to be a significant parameter in the mutability model (Wald $\chi^2 = 2.39$, $p = 0.4954$). This was likely due to the low number of certain motifs present among the Y-STRs studied; for example, 5 of the possible 16 tetranucleotide motifs were absent from the Y-STRs considered here, as were 4 of the 9 trinucleotide motifs, as had been classified by Kelkar et al.[40] Because the majority (66%) of Y-STR loci examined were tetranucleotide repeats, these data were separately analyzed, providing statistically significant differences between the mutation rates of different sequence motifs ($F = 3.29$, $p = 0.0003$), with AAAG having the highest mutation rates (average of 3.57×10^{-3}), followed by AGAT (2.56×10^{-3}), AAAT (9.16×10^{-4}), and AAGG (7.2×10^{-4}).

Male Relative Differentiation by RM Y-STRs

We identified 13 RM Y-STR markers (all with mutation rates $> 1 \times 10^{-2}$): DYF387S1, DYF399S1, DYF403S1, DYF404S1, DYS449, DYS518, DYS526, DYS547, DYS570, DYS576, DYS612, DYS626, and DYS627 (Figure 1; Table S1). Four of these 13 RM Y-STR markers are multicopy systems (DYF387S1 with two copies, DYF399S1 with three copies, DYF403S1 with four copies, DYF404S1 with two copies, and DYS526 with two copies), whereas nine were single-copy Y-STR markers (although six of these markers contained multiple Y-STR loci within the single amplicon, and only two, DYS570 and DYS576, were simple repeats, with only one Y-STR locus, respectively). The 13 RM Y-STRs were combined into a set under the hypothesis that closely related males (even father-son or brother pairs) may be differentiable by Y-STR mutations if RM Y-STRs are combined. In principle, one mutation at one of the 13 RM Y-STRs would be enough for individual differentiation. In order to define a statistical expectation for the RM Y-STR set to differentiate between male relatives and in order to compare their potential with that of the commonly used Yfiler set, we first computed the mutation rate observed for each of the two Y-STR sets by means of a Bayesian approach. The number of mutations observed in each father-son pair for each set of Y-STRs was modeled by means of a Poisson distribution. For the RM Y-STRs, a median mutation rate of 1.97×10^{-2} (95% CI, $1.8 \times 10^{-2} - 2.2 \times 10^{-2}$) of the posterior distribution was estimated, which was 6.5-fold higher than that estimated for Yfiler Y-STRs, with a median rate of 3.0×10^{-3} (95% CI, ranging from 2.39×10^{-3} to 3.72×10^{-3}). Next, we estimated the probability of observing at least one mutation in each of the two Y-STR sets for a given father-son pair, reflecting the minimal criteria for differentiating male relatives. Assuming that all Y-STRs per set were genotyped successfully, and using the posterior estimates of the mutation rate for each set of Y-STR markers, the probability of observing at least one mutation with the RM Y-STR set was 0.1952 (95% CI, 0.177 to 0.21). This value was more than 4 times higher than that estimated for the Yfiler set with 0.047 (95% CI, 0.038 to 0.057). The probability of observing at least one mutation with the RM Y-STR set was statistically significantly higher than for the Yfiler set ($p < 5.0 \times 10^{-07}$). Finally, we empirically tested in samples independently of those used for mutation rate establishment whether the new RM Y-STR set is practically useful

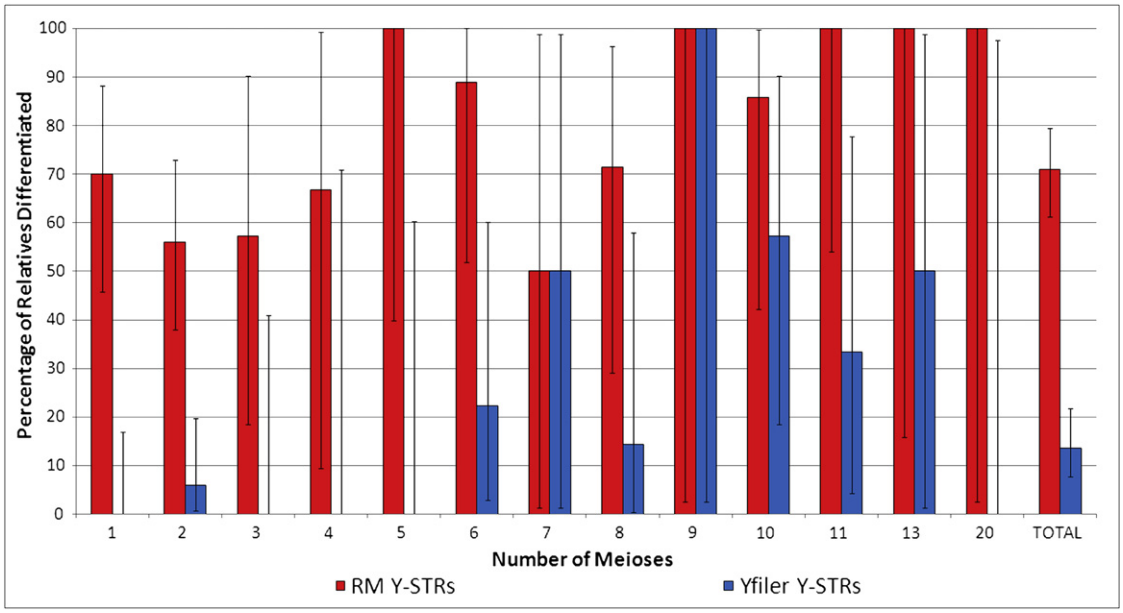

Figure 4. Male Relative Differentiation with Newly Identified 13 RM Y-STRs and Commonly Used 17 Yfiler Y-STRs
Results from differentiating between male relatives from analyzing 103 pairs from 80 male pedigrees, sorted according to the number of generations separating pedigree members, based on 13 RM Y-STRs (in red) and 17 Yfiler Y-STRs (in blue). Error bars represent 95% binomial confidence intervals. Note that these samples are independent from the father-son pairs initially used to establish the Y-STR mutation rates.

for differentiating male relatives. For this, we genotyped an additional 103 male relative pairs from 80 male pedigrees who were related by between 1 and 20 generations within their pedigrees and compared the findings with those we obtained from Yfiler in the same samples. Overall, the RM Y-STR set distinguished 70.9% pairs of male relatives by at least one mutation, reflecting a 5-fold increase in the level of male relative differentiation compared to the Yfiler set with only 13%; notably, the significant difference ($t = 6.389$, $p < 0.0001$) is similar to our statistical expectations from the initial father-son pair analyses (Figure 4; Table S3). Within the pedigrees, the RM Y-STR set distinguished 70% of father-son pairs, 56% of brothers, and 67% of cousins (Figure 4; Table S3). In contrast, the Yfiler set was not able to differentiate any of the father-son pairs or cousins, and it only differentiated 6% of the brothers in this data set (Figure 4; Table S3). Furthermore, all relatives separated by more than 11 generations were differentiable by one or more mutations with the RM Y-STR set, but only 33% were differentiable with the Yfiler set.

Discussion

The mutation rate knowledge we provide here for a large number of Y-STR markers in a reasonably large number of up to 2000 DNA-confirmed father-son pairs extends the range of mutation rates known for Y-STRs, particularly at the upper but also at the lower limits. This study therefore considerably increases the current knowledge of Y-STR mutation rates. The obtained average Y-STR mutation rate of 3.35×10^{-3} across all 186 markers corresponds closely to that found by Kayser et al.[24] and Hohoff et al.[45] for a small number of Y-STRs. However, we noticed a large variation in mutation rates between Y-STR markers, which has not been observed before. This implies that the current approach of applying average mutation rates for Y-STRs is problematic, and only locus-specific knowledge should be used in subsequent applications such as evolutionary or forensic studies. The Y-STR markers included in the present study were ascertained from the literature,[29] mainly from our earlier study that described 167 Y-STR markers for the first time.[29] In this earlier study, markers were found by scanning in silico the human Y chromosome sequence for tandem repeats, with subsequent laboratory testing. No assumptions for Y-STR search were considered except a minimal homogeneous repeat number of 8 and a repeat length equal or greater than 3 (trinucleotide Y-STRs).[29] Consequently, statements about the relation between repeat number and mutability in the present study are influenced by this repeat number threshold, and statements about repeat size and mutability exclude the potential influences of dinucleotides. An additional ascertainment for the present study was that mostly single-copy Y-STRs were used (although some multicopy Y-STRs were added, especially those that showed higher diversity values in previous studies). However, because an independent mutational mechanism of the different copies of a Y-STR marker is assumed, this should not influence the present conclusions on Y-STR mutability. A comparison of the mutation rates for the 17 Yfiler Y-STRs included here (and reported separately eslewhere[28]), with those compiled on the YHRD website (see Web Resources) for

the same markers considering between 4,712 and 21,408 meiotic transfers per marker (excluding data from [28]), showed no significant differences ($t = -0.958$, $p = 0.353$). From this evidence, albeit limited to only 17 Y-STRs, we may conclude that the Y-STR mutation rates estimated here from ~2000 meiotic transfers per marker are reliable, including those for the remaining 169 Y-STR markers, most of which had no mutation data available before.

The wide range of mutation rates observed, and the large amount of sequence data generated, allowed an in-depth investigation into the causes of microsatellite mutation at the molecular level. The model presented partly confirms previous knowledge but also extends it on the influence of specific factors. The length of the repeat was one of the first factors identified as influencing mutation rates of microsatellites,[39,46,47] with a strong inverse relationship observed. Although the repeat length was seen to be a significant factor in the mutability of Y-STRs in our study, the relationship was somewhat obscured by the differences in the numbers of each class of repeat, and in particular the large number of tetranucleotide loci in the set. As has been postulated for several years, the number of repeats present within a given allele has the strongest effect on the probability of a slippage mutation occurring.[24,31,33,47,48] However, in all reports, the only factor tested for influence on mutation rates was the number of repeats in the longest homogenous array. Kayser et al.[29] examined the effect of surrounding repeats on the observed variance of Y-STR loci, but in contrast to results presented here found that the longest homogenous array length was more strongly correlated. However, they used repeat variance as an indirect measure of mutation rate, whereas in the current study we directly applied experimentally derived mutation rate estimates, expected to provide more accurate answers. In addition, only simple repeats are most commonly included in the analyses, because complex repeats can be difficult to classify without extensive sequence data. Because of the large number of complex loci and the extensive sequencing performed, we were able to comprehensively analyze the effect of imperfect repeats and highly complex surrounding sequence on the STR mutation process. This has enabled us to present evidence that the imperfect repeat stretches also play a significant part in the mutability of a given Y-STR. The high estimate of the effect size (78%) underlines the major role that this factor plays in Y-STR mutability, and similar effects are expected for STRs in general. Related to the total number of repeats within an STR locus is the complexity of the repeat motifs. The complexity statistic developed by Kayser et al.[29] incorporates information regarding the number of different motifs in the locus and the number of different blocks of repeats of the same motif. When all the repeats within a locus are included, a significant positive correlation is seen between the complexity and the mutation rate. This is contrary to established theory, because interruptions within repeat arrays have long been thought to decrease mutation rates.[49,50] However, when allele-specific Y-STR mutation rates are compared between matched simple and complex loci (for alleles with 8–16 repeats in the longest homogenous array), no significant difference is found ($p = 0.594$). Even when the Y-STR mutation rates are compared between allele lengths based on the total numbers of repeats, the two repeat classes have similar mutation rates, despite comparing the longest (and thus most mutable) simple repeats with the smallest complex repeats. This would indicate that interruptions do not decrease mutation rates, provided there are sufficient repeats present in the two resulting blocks to maintain a similar level of mutability. Instead, it seems likely that the imperfect repeats surrounding the main array create increased levels of secondary structure within the STR region, increasing the probability of a strand mispairing occurring. Secondary structure, caused by different motifs, was seen to have a large effect on the length distributions of STRs within the human genome, leading to the hypothesis that base stacking within repeats plays a key role in the formation and maintenance of repetitive segments.[51] Although the Y-STRs studied here did not allow us to observe a significant effect of the sequence motif on the mutation rate, the motifs with strong purine:pyrimidine asymmetries (such as AAAG, AGAT, and AAGG) showed higher variance and diversity, suggesting that this may indeed play a role in increasing repeat lengths. Key evidence for the strong effect of the three factors identified to mostly determine Y-STR mutability (repeat length, total repeat number, and total complexity) is seen by examining the 13 RM Y-STRs. The average total repeat number for RM Y-STRs, at 32.8, is more than 2 times greater than for the remaining 173 Y-STRs (15.5). Complexity in RM Y-STRs is increased 2.7-fold, and two population-based estimates, repeat variance and diversity, are also more than doubled in the RM set, compared with the remaining Y-STRs. Thus, it would appear that when optimal levels of the three key parameters are present, the mutability of a given Y-STR can increase by an order of magnitude or more. The elucidation of the features of these rapidly mutating loci, namely small repeat length, high numbers of repeats, and high complexity, may allow the identification of autosomal RM STRs, to complement the RM Y-STRs identified here, for use in downstream applications.

The identification of factors influencing the mutation of STRs on the Y chromosome allows the examination of the various models of microsatellite evolution that have been proposed. The majority of our data is in line with the classical SMM, because we observed 96% of the Y-STR mutations representing single-repeat changes. However, the 4% of the mutations observed as multistep also provide evidence that the SMM does not cover all of Y-STR mutability. The clear bias in mutation rates between allele lengths and the excess of contraction mutations in longer alleles suggests that the modified SMM, as suggested by Xu et al.[34] and others, is necessary to obtain the finite distributions of alleles that we and others have observed.

However, the mutation bias model is not the only theory for explaining microsatellite evolution. Bell and Jurka[52] and Kruglyak et al.[36] proposed point mutations as an additional mechanism to ensure a stationary distribution of alleles. As the repeat length increases, the probability that a point mutation will occur, thus splitting the repeat into two separate blocks, increases. Thus, with the decreased length of the homogenous array, the mutation rate is expected to slow down, preventing infinite growth of the microsatellite. There have been suggestions that the point mutation rate in microsatellites is approximately twice that of nonrepetitive DNA.[38] However, within this study, no point mutations were observed between the sequenced father-son pairs, and only one Y-STR was observed with a variable SNP in the repeat region (DYS624). Thus, it would seem that, although point mutations are important on evolutionary timescales for maintaining repeat lengths (as evidenced by the high complexity of many STRs), they operate considerably more slowly (by an order of magnitude or more) than the slippage bias mechanism.

We also compared the Y-STR mutation rates obtained here to published autosomal mutation rates, aiming to test the possible effect of recombination on STR mutability. Because of the increased mutation rates observed in males,[53] only data from male autosomal meioses[54–56] were used, and only tri-, tetra-, and pentanucleotide repeats were compared. No significant differences between autosomal and Y-STR mutation rates were found ($Z = -0.211$, $p = 0.833$), confirming previous observations that recombination plays a little part in the mutation processes of STR loci.[24,57,58]

With this study, we provide mutation rate estimates for a large number of Y-STR markers. These mutation rate estimates are now available for application in studies that use these Y-STRs, e.g., to address questions of human population and evolutionary history or genealogy or to address questions in the forensic context. The extremely high mutation rates of the RM Y-STRs confer substantial power to differentiate between male relatives. Although all mutation rates reported here were estimated from European populations, diversity values of the RM Y-STRs do not show any significant differences between Europeans and other worldwide groups (unpublished data). Therefore, it may be expected that the extraordinary mutational features of the RM Y-STRs, as described here for Europeans, also hold true for other populations, although mutation rate studies in additional populations will need to be carried out in the future. Commonly used Y-STR sets such as Yfiler, with lower average mutation rates, have shown reduced abilities to reliably differentiate between close male relatives,[12,20] as confirmed here. Although Y-STRs with slower mutation rates are useful for evolutionary and genealogical studies, as well as for paternity testing in deficiency cases with unavailable fathers and male offspring, applications for Y-STRs also exist in which the resolution of Y chromosome haplotyping needs to be finer than tens of generations, such as in forensics, and RM Y-STRs will be beneficial. Hence, the mutation rate knowledge we provide here for a large number of Y-STRs allows the future selection of the most suitable set of Y-STRs, with the most appropriate lineage differentiation level, to be tailored to the application at hand. One such example is forensic identification of males, in which individual conclusions are of vital importance but cannot be achieved with the currently available Y-STR sets because of the low underlying mutation rates. The practical ability of the new RM Y-STR set to differentiate male relatives was aptly shown by the high levels of mutations within the pedigrees, allowing over 70% of relatives to be separated (compared with only 13% with the commonly used Yfiler Y-STRs). It should be noted that, because of multiplex failure and/or sample degradation, we had to exclude 58 additional pairs of relatives from the analysis in which at least one marker failed to amplify, together with no mutation being observed in those markers that were successfully genotyped. Although genotyping failure is not expected to be correlated with mutations, this procedure may have produced a bias. This may explain the comparatively high differentiation rate between fathers and sons in the additional families and pedigrees, which is ~3-fold greater than expected from our simulation analyses based on the data from the initial set of ~2000 father-son pairs (for which different multiplex assays with lower failure rates were applied). However, the high rate of differentiation between male relatives, even with the suboptimal success rate of the multiplex assays used, suggests that this set of RM Y-STRs is highly useful for individualizing male lineages via Y chromosome analysis. In the future, more efficient multiplex assays for the RM Y-STRs introduced here will be developed to take full advantage of the marker properties for male relative differentiation. We see two scenarios for the application of RM Y-STRs in forensic DNA analysis. First, they may be applied to cases with a specific hypothesis about the involvement of related males, such as rape cases with several related perpetrators, in which autosomal STR profiling and conventional Y-STR profiling are usually not informative. Second, and quantitatively more importantly, they may be applied to all forensic cases in which conventional Y-STR (e.g., Yfiler) profiling was applied but did not provide an exclusion constellation. In such cases, the question remains of whether the same man or different but related men were involved; subsequent analysis of the RM Y-STR set we introduce here will provide further evidence for answering this question.

In conclusion, we present mutation rate estimates for the largest number of 186 Y-STRs available to date and with the confidence provided from ~2000 meiotic transfer studies at each marker that we make available for future studies. With the knowledge provided here, it will now be possible for researchers to select a custom set of Y-STRs suitable for various applications, such as slowly mutating Y-STRs for evolutionary studies, medium-mutating Y-STRs

for population history and genealogy studies, and rapidly mutating Y-STRs for microevolutionary studies, for investigating histories of populations that have experienced bottleneck or founder histories, and for forensic applications. The evolutionary mechanisms driving the increased mutation rates at these markers revolve around the repeat length, the total number of repeats present in a locus, and the complexity of the repeat motif and surrounding sequence. Beyond Y-STRs, our findings are expected to be relevant for understanding microsatellite evolution in general. These data have also allowed us to identify 13 Y-STR markers with exceptionally high mutability (termed RM Y-STRs), which provides greatly increased male relative differentiation and will shift forensic Y chromosome analysis from previous male lineage differentiation toward future male individual identification.

Supplemental Data

Supplemental Data include three tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

We would like to thank Damla Arslantunali, Ines Correia Rosa, Paul Dekkers, Sevgi Deniz, Kristiaan van der Gaag, Anita van den Heuvel, Diana van den Heuvel, Cissi Jonsson-Glans, Tomas Petten, Arwin Ralf, and Melike Yüksel for expert technical assistance. Pieter van Oers (Life Technologies) is thanked for infrastructural support. This work was supported by funds from the Netherlands Forensic Institute to M.K. and was additionally supported by a grant from the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research within the framework of the Forensic Genomics Consortium Netherlands to M.K. and P.d.K. R.P. is supported by a Polish Ministry of Science and Higher Education grant N404 032 31/1892. R.F. and M.R.F. are employees and shareholders of Life Technologies. This work was supported in part by Life Technologies.

Received: July 5, 2010
Revised: August 2, 2010
Accepted: August 13, 2010
Published online: September 2, 2010

Web Resources

The URLs for data presented herein are as follows:

The R Project for Statistical Computing, http://www.r-project.org/
YHRD: Y-STR Haplotype Reference Database, www.yhrd.org/

References

1. Underhill, P.A., and Kivisild, T. (2007). Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. Annu. Rev. Genet. *41*, 539–564.
2. Kayser, M. (2007). Uni-parental markers in human identity testing including forensic DNA analysis. Biotechniques *43*, 3042.
3. Karafet, T.M., Mendez, F.L., Meilerman, M.B., Underhill, P.A., Zegura, S.L., and Hammer, M.F. (2008). New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res. *18*, 830–838.
4. Jobling, M.A., and Tyler-Smith, C. (2003). The human Y chromosome: An evolutionary marker comes of age. Nat. Rev. Genet. *4*, 598–612.
5. King, T.E., and Jobling, M.A. (2009). Founders, Drift and Infidelity: The relationship between Y chromosome diversity and patrilineal surnames. Mol. Biol. Evol. *26*, 1093–1102.
6. Oota, H., Settheetham-Ishida, W., Tiwawech, D., Ishida, T., and Stoneking, M. (2001). Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. Nat. Genet. *29*, 20–21.
7. Seielstad, M.T., Minch, E., and Cavalli-Sforza, L.L. (1998). Genetic evidence for a higher female migration rate in humans. Nat. Genet. *20*, 278–280.
8. Hammer, M.F., and Zegurea, S.L. (2002). The human Y chromosome haplogroup tree: Nomenclature and phylogeography of its major divisions. Annu. Rev. Anthropol. *31*, 303–321.
9. Kayser, M., Brauer, S., Weiss, G., Underhill, P.A., Roewer, L., Schiefenhövel, W., and Stoneking, M. (2000). Melanesian origin of Polynesian Y chromosomes. Curr. Biol. *10*, 1237–1246.
10. Underhill, P.A., Shen, P., Lin, A.A., Jin, L., Passarino, G., Yang, W.H., Kauffman, E., Bonné-Tamir, B., Bertranpetit, J., Francalacci, P., et al. (2000). Y chromosome sequence variation and the history of human populations. Nat. Genet. *26*, 358–361.
11. Shi, W., Ayub, Q., Vermeulen, M., Shao, R.G., Zuniga, S., van der Gaag, K., de Knijff, P., Kayser, M., Xue, Y., and Tyler-Smith, C. (2010). A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations. Mol. Biol. Evol. *27*, 385–393.
12. Kayser, M., Vermeulen, M., Knoblauch, H., Schuster, H., Krawczak, M., and Roewer, L. (2007). Relating two deep-rooted pedigrees from Central Germany by high-resolution Y-STR haplotyping. Forensic Sci. Int.; Genet. *1*, 125–128.
13. Kayser, M., Caglià, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Heidorn, F., Herrmann, S., Herzog, B., Hidding, M., et al. (1997). Evaluation of Y-chromosomal STRs: A multicenter study. Int. J. Legal Med. *110*, 125–133, 141–149.
14. Dettlaff-Kakol, A., and Pawlowski, R. (2002). First Polish DNA "manhunt"—an application of Y-chromosome STRs. Int. J. Legal Med. *116*, 289–291.
15. Roewer, L. (2009). Y chromosome STR typing in crime casework. Forensic Sci. Med. Pathol. *5*, 77–84.
16. Stumpf, M.P.H., and Goldstein, D.B. (2001). Genealogical and evolutionary inference with the human Y chromosome. Science *291*, 1738–1742.
17. Novelletto, A. (2007). Y chromosome variation in Europe: Continental and local processes in the formation of the extant gene pool. Ann. Hum. Biol. *34*, 139–172.
18. Balaresque, P., Bowden, G.R., Adams, S.M., Leung, H.Y., King, T.E., Rosser, Z.H., Goodwin, J., Moisan, J.P., Richard, C., Millward, A., et al. (2010). A predominantly neolithic origin for European paternal lineages. PLoS Biol. *8*, e1000285.
19. Kayser, M., and Sajantila, A. (2001). Mutations at Y-STR loci: Implications for paternity testing and forensic analysis. Forensic Sci. Int. *118*, 116–121.
20. Heyer, E., Puymirat, J., Dieltjes, P., Bakker, E., and de Knijff, P. (1997). Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum. Mol. Genet. *6*, 799–803.

21. Pollin, T.I., McBride, D.J., Agarwala, R., Schäffer, A.A., Shuldiner, A.R., Mitchell, B.D., and O'Connell, J.R. (2008). Investigations of the Y chromosome, male founder structure and YSTR mutation rates in the Old Order Amish. Hum. Hered. *65*, 91–104.
22. Vermeulen, M., Wollstein, A., van der Gaag, K., Lao, O., Xue, Y., Wang, Q., Roewer, L., Knoblauch, H., Tyler-Smith, C., de Knijff, P., and Kayser, M. (2009). Improving global and regional resolution of male lineage differentiation by simple single-copy Y-chromosomal short tandem repeat polymorphisms. Forensic Sci. Int. Genet. *3*, 205–213.
23. Zhivotovsky, L.A., Underhill, P.A., Cinnioğlu, C., Kayser, M., Morar, B., Kivisild, T., Scozzari, R., Cruciani, F., Destro-Bisol, G., Spedini, G., et al. (2004). The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am. J. Hum. Genet. *74*, 50–61.
24. Kayser, M., Roewer, L., Hedman, M., Henke, L., Henke, J., Brauer, S., Krüger, C., Krawczak, M., Nagy, M., Dobosz, T., et al. (2000). Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am. J. Hum. Genet. *67*, 1526–1543.
25. Dupuy, B.M., Stenersen, M., Flønes, A.G., Egeland, T., and Olaisen, B. (2004). Y-chromosomal microsatellite mutation rates: Differences in mutation rate between and within loci. Hum. Mutat. *23*, 117–124.
26. Gusmão, L., Sánchez-Diz, P., Calafell, F., Martín, P., Alonso, C.A., Alvarez-Fernández, F., Alves, C., Borjas-Fajardo, L., Bozzo, W.R., Bravo, M.L., et al. (2005). Mutation rates at Y chromosome specific microsatellites. Hum. Mutat. *26*, 520–528.
27. Ge, J., Budowle, B., Aranda, X.G., Planz, J.V., Eisenberg, A.J., and Chakraborty, R. (2009). Mutation rates at Y chromosome short tandem repeats in Texas populations. Forensic Sci. Int. Genet. *3*, 179–184.
28. Goedbloed, M., Vermeulen, M., Fang, R.N., Lembring, M., Wollstein, A., Ballantyne, K., Lao, O., Brauer, S., Krüger, C., Roewer, L., et al. (2009). Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFlSTR Yfiler PCR amplification kit. Int. J. Legal Med. *123*, 471–482.
29. Kayser, M., Kittler, R., Erler, A., Hedman, M., Lee, A.C., Mohyuddin, A., Mehdi, S.Q., Rosser, Z., Stoneking, M., Jobling, M.A., et al. (2004). A comprehensive survey of human Y-chromosomal microsatellites. Am. J. Hum. Genet. *74*, 1183–1197.
30. Ota, T., and Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. *22*, 201–204.
31. Lai, Y., and Sun, F. (2003). The relationship between microsatellite slippage mutation rate and the number of repeat units. Mol. Biol. Evol. *20*, 2123–2131.
32. Ellegren, H. (2000). Heterogeneous mutation processes in human microsatellite DNA sequences. Nat. Genet. *24*, 400–402.
33. Xu, X., Peng, M., Fang, Z., and Xu, X. (2000). The direction of microsatellite mutations is dependent upon allele length. Nat. Genet. *24*, 396–399.
34. Amos, W., Sawcer, S.J., Feakes, R.W., and Rubinsztein, D.C. (1996). Microsatellites show mutational bias and heterozygote instability. Nat. Genet. *13*, 390–391.
35. Kruglyak, S., Durrett, R.T., Schug, M.D., and Aquadro, C.F. (1998). Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. USA *95*, 10774–10778.
36. Whittaker, J.C., Harbord, R.M., Boxall, N., Mackay, I., Dawson, G., and Sibly, R.M. (2003). Likelihood-based estimation of microsatellite mutation rates. Genetics *164*, 781–787.
37. Pumpernik, D., Oblak, B., and Borštnik, B. (2008). Replication slippage versus point mutation rates in short tandem repeats of the human genome. Mol. Genet. Genomics *279*, 53–61.
38. Eckert, K.A., and Hile, S.E. (2009). Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. Mol. Carcinog. *48*, 379–388.
39. Pemberton, T.J., Sandefur, C.I., Jakobsson, M., and Rosenberg, N.A. (2009). Sequence determinants of human microsatellite variability. BMC Genomics *10*, 612.
40. Kelkar, Y.D., Tyekucheva, S., Chiaromonte, F., and Makova, K.D. (2008). The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res. *18*, 30–38.
41. Hedman, M., Neuvonen, A.M., Sajantila, A., and Palo, J.U. (2010). Dissecting the Finnish male uniformity: The value of additional Y-STR loci. Forensic Sci. Int. Genet., in press. Published online April 28, 2010.
42. Hanson, E.K., Berdos, P.N., and Ballantyne, J. (2006). Testing and evaluation of 43 "noncore" Y chromosome markers for forensic casework applications. J. Forensic Sci. *51*, 1298–1314.
43. Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003). Bayesian Data Analysis, Second Edition (Boca Raton, FL: Chapman & Hall).
44. Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modeling framework: Concepts, structure and extensibility. Stat. Comput. *10*, 325–337.
45. Hohoff, C., Dewa, K., Sibbing, U., Hoppe, K., Forster, P., and Brinkmann, B. (2007). Y-chromosomal microsatellite mutation rates in a population sample from northwestern Germany. Int. J. Legal Med. *121*, 359–363.
46. Chakraborty, R., Kimmel, M., Stivers, D.N., Davison, L.J., and Deka, R. (1997). Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc. Natl. Acad. Sci. USA *94*, 1041–1046.
47. Webster, M.T., Smith, N.G.C., and Ellegren, H. (2002). Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. Proc. Natl. Acad. Sci. USA *99*, 8748–8753.
48. Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., and Rolf, B. (1998). Mutation rate in human microsatellites: Influence of the structure and length of the tandem repeat. Am. J. Hum. Genet. *62*, 1408–1415.
49. Ellegren, H. (2004). Microsatellites: Simple sequences with complex evolution. Nat. Rev. Genet. *5*, 435–445.
50. Sainudiin, R., Durrett, R.T., Aquadro, C.F., and Nielsen, R. (2004). Microsatellite mutation models: Insights from a comparison of humans and chimpanzees. Genetics *168*, 383–395.
51. Bacolla, A., Larson, J.E., Collins, J.R., Li, J., Milosavljevic, A., Stenson, P.D., Cooper, D.N., and Wells, R.D. (2008). Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties. Genome Res. *18*, 1545–1553.

52. Bell, G.I., and Jurka, J. (1997). The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single-step mutation process. J. Mol. Evol. *44*, 414–421.
53. Makova, K.D., and Li, W.-H. (2002). Strong male-driven evolution of DNA sequences in humans and apes. Nature *416*, 624–626.
54. Leopoldino, A.M., and Pena, S.D.J. (2002). The mutational spectrum of human autosomal tetranucleotide microsatellites. Hum. Mutat. *21*, 71–79.
55. Henke, J., and Henke, L. (1999). Mutation rate in human microsatellites. Am. J. Hum. Genet. *64*, 1473–1474.
56. American Association of Blood Banks. (2006). Annual Report Summary, Parentage Testing Standards Program. http://www.dnatestingcentre.com/Reports/06AABB_summary.pdf
57. Huang, Q.-Y., Xu, F.-H., Shen, H., Deng, H.-Y., Liu, Y.-J., Liu, Y.-Z., Li, J.-L., Recker, R.R., and Deng, H.-W. (2002). Mutation patterns at dinucleotide microsatellite loci in humans. Am. J. Hum. Genet. *70*, 625–634.
58. Morral, N., Nunes, V., Casals, T., and Estivill, X. (1991). CA/GT microsatellite alleles within the cystic fibrosis transmembrane conductance regulator (CFTR) gene are not generated by unequal crossingover. Genomics *10*, 692–698.