Empirical Evaluation Reveals Best Fit of a Logistic Mutation Model for Human Y-Chromosomal Microsatellites

Arne Jochens,^{*,†} Amke Caliebe,^{*} Uwe Rösler,[†] and Michael Krawczak^{*}

^{*}Institut für Medizinische Informatik und Statistik, Christian-Albrechts-Universität zu Kiel, 24105 Kiel, Germany, and

[†]Mathematisches Seminar, Christian-Albrechts-Universität zu Kiel, 24098 Kiel, Germany

ABSTRACT The rate of microsatellite mutation is dependent upon both the allele length and the repeat motif, but the exact nature of this relationship is still unknown. We analyzed data on the inheritance of human Y-chromosomal microsatellites in father–son duos, taken from 24 published reports and comprising 15,285 directly observable meioses. At the six microsatellites analyzed (DYS19, DYS389I, DYS390, DYS391, DYS392, and DYS393), a total of 162 mutations were observed. For each locus, we employed a maximum-likelihood approach to evaluate one of several single-step mutation models on the basis of the data. For five of the six loci considered, a novel logistic mutation model was found to provide the best fit according to Akaike's information criterion. This implies that the mutation probability at the loci increases (nonlinearly) with allele length at a rate that differs between upward and downward mutations. For DYS392, the best fit was provided by a linear model in which upward and downward mutation probabilities increase equally with allele length. This is the first study to empirically compare different microsatellite mutation models in a locus-specific fashion.

"MICROSATELLITES", also known as "short tandem repeats" (STRs), are stretches of DNA in which a short piece of sequence (1–6 bp, the repeat motif) is tandemly repeated a number of times. Microsatellites are abundant throughout the human genome (Lander *et al.* 2001). They have high mutation rates, which renders them useful for a number of practical applications, including genetic epidemiology (Jobling and Tyler-Smith 2003), gene mapping (Weissenbach 1993), studies of population history (Jobling and Tyler-Smith 2003), genetic fingerprinting (Krawczak and Schmidtke 1998), kinship testing (Weir *et al.* 2006), and genetic ancestry analysis (Shriver and Kittles 2004). Microsatellites also may have played a major role in genome evolution (Kashi and King 2006). Microsatellite mutations typically comprise the gain or loss of 1 repeat unit, but larger changes of the repeat number in a single meiosis also have been observed (see below).

Various statistical models have been proposed for the microsatellite mutation process (Calabrese and Sainudiin 2005), but no model has yet been identified as representing the single best, thereby testifying to the complexity of the mutation process. In any case, it is well established that the mutation rate of microsatellites (i) depends upon the respective repeat number (Xu *et al.* 2000; Lai and Sun 2003; Kelkar *et al.* 2008) and (ii) varies greatly across loci, which is partly due to the variable nature of the repeat motif itself (Kelkar *et al.* 2008).

The stepwise mutation model (SMM) was the first one used to describe microsatellite evolution. The SMM implies that each mutation adds or deletes 1 repeat unit at a time with a symmetric mutation rate that is independent of the repeat number (Ohta and Kimura 1973). The SMM was used in simulation-based attempts to explain the observed population patterns of microsatellite allele frequencies (Shriver *et al.* 1993; Valdes *et al.* 1993). Since then, various modifications of the SMM have been proposed. For example, the two-phase model (Di Rienzo *et al.* 1994) combines the SMM with geometrically distributed mutational changes of more than 1 repeat unit. This model was itself generalized (Garza *et al.* 1995), allowing for any prespecified distribution of the repeat number change. Other approaches included lower

Copyright © 2011 by the Genetics Society of America

doi: 10.1534/genetics.111.132308

Manuscript received July 6, 2011; accepted for publication September 23, 2011

[†]Corresponding author: Institut für Medizinische Informatik und Statistik, Christian-Albrechts-Universität zu Kiel, Haus 31, Arnold-Heller-Str. 3, 24105 Kiel, Germany.

E-mail: jochens@medinfo.uni-kiel.de

and upper bounds for the allele length (Nauta and Weissing 1996; Feldman *et al.* 1997).

Several microsatellite mutation models have been proposed to account for the relationship between the mutation rate and the respective allele length (Falush and Iwasa 1999). More restricted versions of this approach entail the idea that only the difference in repeat number before and after mutation, but not the mutation rate itself, depends upon the allele length (Garza *et al.* 1995; Kimmel *et al.* 1996). Kruglyak, Durrett, and co-workers (Kruglyak *et al.* 1998; Durrett and Kruglyak 1999) combined the SMM with the possibility of (rare) point mutations that reduce the repeat number. This model, later expanded (Calabrese *et al.* 2001; Dieringer and Schlötterer 2003), is motivated by the idea that a point mutation divides a microsatellite into two parts, only one of which is kept track of.

Finally, Whittaker *et al.* (2003) proposed an exponential microsatellite mutation model with parameters α_u , α_d , γ_u , γ_d , λ_u , and λ_d from an unspecified parameter space. Their transition probabilities (*cf.* *Materials and Methods*) from paternal allele i to offspring allele j were as follows:

$$p_{ij} = \begin{cases} \gamma_u \cdot \exp(\alpha_u \cdot i - \lambda_u \cdot (j - i)) & \text{if } j > i, \\ \gamma_d \cdot \exp(\alpha_d \cdot i - \lambda_d \cdot (i - j)) & \text{if } j < i, \\ 1 - \sum_{k \neq i} p_{ik} & \text{if } j = i. \end{cases}$$

However, in mathematical terms this model is not well defined because, for fixed parameters with α_u , $\gamma_u > 0$, $p_{i,i+1}$ may exceed 1 for large i , and $p_{i,i+1}$ even approaches infinity as $i \rightarrow \infty$. Nevertheless, the model can be remedied if written as an instantaneous rate matrix over a finite state space as in Calabrese and Sainudiin (2005).

In the present study, we compared different models of microsatellite mutation, taking into account the locus dependency of the mutation process by evaluating each model separately for each locus. We chose to confine our study to microsatellites from the male-specific region of the human Y chromosome (Y-STRs) for which mutations can be inferred directly from genotype data on father–son duos. A large amount of such data have become available for opportunistic use in scientific analyses because large numbers of father–son pairs are regularly genotyped in paternity cases. Moreover, these data allow us to directly infer the one-step probability matrix of the discrete-time Markov chain model scaled in generations, as opposed to the continuous-time Markov process usually applied to evolutionary genomic data.

Materials and Methods

Y-STRs

In forensics as well as for other practical applications, a few commercially available, PCR-based kits are widely used for microsatellite genotyping (Mayntz-Press and Ballantyne 2007). The Y chromosome is of particular interest in foren-

Table 1 Y-STR loci

Name	μ [10^{-3}]	Alleles	Repeat motif
DYS19	2.3	14 (9–19)	TAGA
DYS389I	2.5	13 (9–17)	TCTG/TCTA
DYS390	2.1	24 (17–29)	TCTG/TCTA
DYS391	2.6	10 (5–16)	TCTA
DYS392	0.4	11 (4–20)	TAT
DYS393	1.0	13 (7–18)	AGAT

Characteristics of six Y-STR loci considered in the present study are shown. μ is the relative frequency of observed mutations across all paternal alleles for each locus. The “Alleles” column contains the most frequently observed repeat number, and the range is given in parentheses. Data are from the Y Chromosome Haplotype Reference Database (Willuweit and Roewer 2007), Release 33.

sics because it allows easy discrimination between male and female contributions to traces and trace mixtures. This has led to a situation where a few Y-STRs are routinely genotyped in a vast number of individuals (Roewer 2009). Table 1 lists the six Y-STRs that have been included in the present study, along with their most important molecular characteristics. For labeling Y-STR loci and alleles, we employed the nomenclature recommended by the International Society of Forensic Genetics (Gusmão *et al.* 2006). In particular, the allele designations refer to the actual number of repeats rather than the crude PCR fragment length.

Mutation data

Table 2 lists 24 published reports of directly observed Y-STR mutations (or the lack thereof) in father–son pairs, which formed the basis of our present study. Taken together, these reports comprised 15,285 father–son pairs in whom 162 mutations were observed across the six Y-STRs considered. Paternity was confirmed in all cases by independent genotyping of autosomal loci. All studies were original reports so that there should be no overlap between the samples included. Most of the recent reports included the population frequencies of the paternal alleles along with the mutation data, as was recommended by the International Society of Forensic Genetics in 2006 (Gusmão *et al.* 2006). This information is required for fitting mutation models with allele-dependent mutation rates. For studies that did not provide paternal allele frequencies, we tried to approximate these as follows: For four studies (Kayser *et al.* 2000; Budowle *et al.* 2005; Ge *et al.* 2009; Goedbloed *et al.* 2009), estimates were equated to the respective allele frequencies as reported for the relevant subpopulations in the Y Chromosome Haplotype Reference Database (Willuweit and Roewer 2007), Release 31. For the remaining studies, we adopted the allele frequencies reported in the studies themselves, although these estimates included not only fathers, but also unrelated individuals.

Mutation models

The microsatellite mutation models considered here share a number of characteristics. Thus, alleles are consistently represented by an integer corresponding to the number of

Table 2 Reports on father–son duo Y-STR data

Reference	Sample origins	<i>N</i>	<i>n</i>
Bianchi <i>et al.</i> (1998)	United States, France, Venezuela	249	0
Lessig and Edelmann (1998)	Leipzig (Germany)	41	1
Pestoni <i>et al.</i> (1999)	Galicia (Spain)	35	0
Kayser <i>et al.</i> (2000)	Germany, Poland	996	9
Dupuy <i>et al.</i> (2001)	Norway	150	3
Tsai <i>et al.</i> (2002)	Taiwan	109	1
Dupuy <i>et al.</i> (2004)	Norway	1,766	24
Kurihara <i>et al.</i> (2004)	Japan	161	2
Ballard <i>et al.</i> (2005)	United Kingdom	248	4
Berger <i>et al.</i> (2005)	Tyrol (Austria)	70	1
Budowle <i>et al.</i> (2005)	North America	692	5
De Souza Góes <i>et al.</i> (2005)	Rio de Janeiro (Brazil)	119	4
Gusmão <i>et al.</i> (2005)	South America, Portugal, Spain	2,816	23
Turrina <i>et al.</i> (2006)	Northeast Italy	50	1
Domingues <i>et al.</i> (2007)	Rio de Janeiro (Brazil)	135	1
Hohoff <i>et al.</i> (2007)	Westphalia (Germany)	1,029	12
Lee <i>et al.</i> (2007)	Korea	369	5
Pontes <i>et al.</i> (2007)	North Portugal	45	0
Decker <i>et al.</i> (2008)	Ohio	389	7
Padilla-Gutiérrez <i>et al.</i> (2008)	Mexico	218	0
Sánchez-Diz <i>et al.</i> (2008)	Argentina, Brazil, Portugal	701	9
Soares-Vieira <i>et al.</i> (2008)	São Paulo (Brazil)	222	6
Ge <i>et al.</i> (2009)	Texas	2,913	17
Goedbloed <i>et al.</i> (2009)	Germany, Poland	1,762	27
Sum		15,285	162

Reports on directly observed Y-STR mutations in father–son duos are shown. *N* denotes the number of observed father–son duos included, and *n* is the number of observed mutations across all six loci considered in this study.

repeats. In a father-to-son transmission, an allele may change according to a probability distribution that depends only upon the paternal allele. Mathematically, this means that each model entails a time-homogeneous Markov chain $(X_n)_{n \in \mathbb{Z}}$ with state space $S = \mathbb{Z}$, the set of all integers. Each mutation model is thus defined by specifying, for all $i, j \in \mathbb{Z}$, transition probabilities p_{ij} from paternal allele i to offspring allele j . For values of j not explicitly addressed in the definitions given below, we assume $p_{ij} = 0$. Since the data considered here contained only three mutations that changed the paternal allele by more than 1 repeat unit (Table 3), we confined our analyses to single-step models, characterized by $p_{ij} = 0$ for all $i, j \in \mathbb{Z}$ with $|j - i| > 1$.

Stepwise mutation model: The most popular model of microsatellite mutation is the SMM, originally proposed by Ohta and Kimura (1973), although in a slightly different context. The parameters of the SMM are $0 < \mu_u, \mu_d < 0.5$, and the transition probabilities are

$$p_{ij} = \begin{cases} \mu_u & \text{if } j = i + 1, \\ \mu_d & \text{if } j = i - 1, \\ 1 - \mu_u - \mu_d & \text{if } j = i. \end{cases}$$

Parameters μ_u and μ_d correspond to the upward and downward mutation rates, respectively. The symmetric SMM is a special case in which $\mu_u = \mu_d = \mu$.

Linear model: To be able to properly define the linear model (Kruglyak *et al.* 1998), we must restrict the state

space to $S = \{1, 2, \dots, k\}$, with $k \in \mathbb{N}$ fixed. The linear model has parameters $\nu_u, \nu_d \in [0, 1/(2k)]$, and the transition probabilities are

$$p_{ij} = \begin{cases} i \cdot \nu_u & \text{if } j = i + 1, \\ i \cdot \nu_d & \text{if } j = i - 1, \\ 1 - i \cdot (\nu_u + \nu_d) & \text{if } j = i, \end{cases}$$

for all $i \in S \setminus \{1, k\}, j \in S$. Furthermore, $p_{12} = \nu_u, p_{11} = 1 - \nu_u, p_{k,k-1} = k \cdot \nu_d$, and $p_{kk} = 1 - k \cdot \nu_d$. The symmetric linear model is a special case in which $\nu_u = \nu_d = \nu$.

Logistic model: Here, we newly introduce the logistic model, defined on state space $S = \mathbb{Z}$. To our knowledge, this mutation model has not been evaluated yet, neither theoretically nor empirically. In its most general form, the logistic model has parameters $\alpha_u, \alpha_d \in \mathbb{R}, \beta_u, \beta_d \in \mathbb{R}^+$, and $\gamma_u, \gamma_d \in (0, 0.5)$ and is characterized by transition probabilities

$$p_{ij} = \begin{cases} \frac{\gamma_u}{1 + \exp(\alpha_u \cdot (\beta_u - i))} & \text{if } j = i + 1, \\ \frac{\gamma_d}{1 + \exp(\alpha_d \cdot (\beta_d - i))} & \text{if } j = i - 1, \\ 1 - \frac{\gamma_u}{1 + \exp(\alpha_u \cdot (\beta_u - i))} - \frac{\gamma_d}{1 + \exp(\alpha_d \cdot (\beta_d - i))} & \text{if } j = i. \end{cases}$$

Thus, α_u, β_u , and γ_u together with the repeat number determine the probability of a gain of 1 repeat unit. Similarly,

Table 3 Observed Y-STR mutations

Allele	DYS19			DYS389I			DYS390			
	-1	0	+1	-1	0	+1	-1	0	+1	+2
8					5					
9					816	1				
10					1123	1				
11		2		2	502					
12		26		2	2126	3				
13		1438		3	6712	7				
14		7017	10	9	2093	2				
15		3847	3	3	45					
16	5	1887	3							
17	7	954	4							
18	2	11								
19		3								
20							21			
21							1055	1		
22							1784		1	
23							6	3691	2	
24							3	5438	10	
25							5	2495	3	
26								203	1	
27								13		

Allele	DYS391					DYS392			DYS393		
	-2	-1	0	+1	+2	-1	0	+1	-1	0	+1
5			1								
6			5								
7			3					1			
8			29					5			
9			526					10			4
10			8446	11				90			12
11		14	5307	12	1	2	6813				40
12	1	3	215				870			1468	
13			20				5356	3	2	9490	5
14			1			1	1128	1	2	1699	3
15							164		1	426	
16							100		1	21	
17							4			4	
18							1				

Numbers of Y-STR mutations observed in 24 studies (cf. Table 2) of father-son duos are shown. "Allele": repeat number of the paternal allele. "-1" and "+1" indicate loss or gain, respectively, of 1 repeat unit in the son. "-2" and "+2" indicate losses or gains of 2 repeat units in one generation. "0": no mutation.

α_d , β_d , and γ_d determine the probability of a downward mutation. Specifically, the α -values determine the increase in mutation rate with increasing allele length while the γ -values can be interpreted as maximum mutation probabilities (mutation probabilities for very long alleles). More formally, $\lim_{i \rightarrow \infty} p_{i,i+1} = \gamma_u$ if $\alpha_u > 0$, and $\lim_{i \rightarrow \infty} p_{i,i-1} = \gamma_d$ if $\alpha_d > 0$. Parameter β can be thought of as the repeat number that corresponds to an intermediate mutation probability $\frac{1}{2}\gamma$.

Statistical approach

Similar to Whittaker *et al.* (2003), we employed a maximum-likelihood approach (Pawitan 2001) to estimate the parameters of a given mutation model. Assuming pairwise independence of the father-to-son transmissions, the likeli-

hood L of a fixed parameter vector θ equals the product of the respective transition probabilities, taken over all observed father-son duos. Thus

$$L(\theta) = \prod_{i=1}^N p_i(\theta), \quad (1)$$

where N is the number of observed father-son duos and $p_i(\theta)$ is the respective transition probability (with parameter vector θ) for the i th father-son duo. Note that Equation 1 includes all observed father-son duos, not only those in whom a mutation occurred, because transition probabilities include the probability of the Markov chain staying in the current state. This is why paternal allele frequencies are required for alleles not involved in any observed mutation.

Since the likelihoods in Equation 1 are typically very small, it is convenient to use logarithms; i.e.,

$$\mathcal{L}(\theta) = \log(L(\theta)) = \sum_{i=1}^N \log(p_i(\theta)). \quad (2)$$

For each of the six Y-STRs and for each model considered, we maximized the log-likelihood function $\mathcal{L}$ using numerical optimization (Nocedal and Wright 2006) as implemented in the *R* environment (Ihaka and Gentleman 1996), particularly the *nlminb* function from the *stats* package.

To allow comparison of unnested mutation models that have different numbers of parameters, we used Akaike's information criterion (AIC), computed as

$$\text{AIC} = -2 \cdot \max_{\theta} \{\mathcal{L}(\theta)\} + 2k, \quad (3)$$

where k is the number of parameters in the mutation model of interest (Akaike 1973; Burnham and Anderson 2002). The smaller the AIC value is for a given Y-STR, the better the model fit. In the case of nested models, any additional parameter has to improve the model fit by at least 1 log-likelihood unit to reduce the AIC value. As a general rule of thumb, all models with an AIC value less than 2 units above the smallest AIC value warrant further attention (Burnham and Anderson 2002).

Results

Observed mutations

The observed distribution of mutations (Table 3) reveals a consistent pattern of upward bias for shorter alleles and downward bias for longer alleles. The only apparent exception is DYS392, but since only seven mutations were observed at this locus, the observed pattern still seems inconclusive.

All observed mutations were single step except for a -2 mutation at DYS391 and one +2 mutation each at DYS390 and DYS391. These mutations were counted as -1 and +1, respectively, for the purpose of model fitting.

Figure 1 Allele-specific mutation frequencies observed at six Y-STR loci. In each panel, top bars depict upward mutations while bottom bars correspond to downward mutations. Where observed frequencies are larger than zero, 95% confidence intervals are plotted as error bars. Note that the depicted frequency range ends at 0.01, which results in the largest downward mutation bar being clipped for five loci. In these cases, the true height of the bar is given inside the bar.

Model fitting

A straightforward approach to estimate locus- and allele-specific mutation probabilities would be by way of observed relative frequencies of mutations (Figure 1). This simplistic assessment already reveals an asymmetry between upward and downward mutations. In particular, long alleles appear to have high downward mutation probabilities. However, since allele-specific probability estimates are not reliable for alleles with few or no observed mutations, a model-based approach appeared warranted. Moreover, mathematical models potentially yield additional insights into the biological nature of the microsatellite mutation process. We used a maximum-likelihood approach to evaluate several mutation models using the Y-STR data summarized in Table

3. The resulting parameter estimates for some of these models are given in Table 4.

Inspection of the model- and locus-specific maximum log-likelihoods and AIC values (Table 5) reveals a remarkable consistency across the first four loci (DYS19, DYS389I, DYS390, and DYS391). For each of these, the symmetric SMM was found to be only a crude approximation, and the asymmetric SMM was not significantly better. The linear model turned out to be superior to the SMM although allowance for asymmetry again did not seem to improve the model fit substantially. The logistic model provided the best fit to the available data for all four loci, and it was found to be substantially better than both the SMM and the linear model. Within the logistic models, symmetry with respect to

Table 4 Parameter estimates

Model	Parameter	Locus-specific estimates					
		DYS19	DYS389I	DYS390	DYS391	DYS392	DYS393
1	μ	1.14×10^{-3}	1.22×10^{-3}	1.07×10^{-3}	1.43×10^{-3}	2.22×10^{-4}	5.41×10^{-4}
2	μ_u	1.31×10^{-3}	1.04×10^{-3}	1.22×10^{-3}	1.64×10^{-3}	2.75×10^{-4}	6.07×10^{-4}
	μ_d	9.20×10^{-4}	1.41×10^{-3}	9.50×10^{-4}	1.23×10^{-3}	2.06×10^{-4}	4.55×10^{-4}
3	ν	7.65×10^{-5}	9.86×10^{-5}	4.62×10^{-5}	1.39×10^{-5}	1.99×10^{-5}	4.06×10^{-5}
4	ν_u	9.00×10^{-5}	8.37×10^{-5}	5.20×10^{-5}	1.59×10^{-5}	2.27×10^{-5}	4.64×10^{-5}
	ν_d	6.30×10^{-5}	1.14×10^{-4}	4.05×10^{-5}	1.19×10^{-5}	1.70×10^{-5}	3.48×10^{-5}
5	α	-0.886	-0.453	-0.766	-1.65	-0.701	-1.13
	β	22.0	24.7	31.9	11.7	13.4	16.7
	γ	0.500	0.278	0.500	0.0115	7.87×10^{-4}	0.0228
8	α_u	0.374	0.0183	0.323	0.592	58.3	0.611
	α_d	2.08	1.18	1.57	17.7	3.47×10^{-3}	1.57
	β_u	27.4	349	37.0	14.1	13.0	20.8
	β_d	18.3	18.0	27.2	11.1	270	16.7
	γ	0.149	0.500	0.0903	0.0163	7.13×10^{-4}	0.0641

Locus-specific maximum-likelihood estimates of the parameters of selected single-step microsatellite mutation models are shown. Models were fitted to the data of Table 3. For model numbers, see Table 5.

γ seems to be acceptable whereas α and β apparently need to be asymmetric. Thus, the newly introduced logistic model with parameters α_u , α_d , β_u , β_d , and γ turned out to provide the best fit according to the AIC, closely followed by the full logistic model. All other model fits were significantly worse.

For DYS392, the symmetric linear model yielded the smallest AIC, but this should be interpreted with caution because only few mutations were observed at this locus (Table

3). For DYS393, the logistic model with symmetrical α -, β -, and γ -values was the best fit according to AIC, but other versions of the logistic model yielded very similar AIC values.

A plot of three models fitted to the DYS19 data (Figure 2) serves to illustrate further that the novel logistic model is appropriate to explain the respective mutation data. The corresponding plots for the other five loci look similar and are therefore not shown.

Table 5 Maximum-likelihood and AIC values

Model	Parameters	DYS19 (N = 15,219)		DYS389I (N = 13,455)		DYS390 (N = 14,732)	
		ΔML	ΔAIC	ΔML	ΔAIC	ΔML	ΔAIC
1	SMM	μ	0	0	0	0	0
2	Linear	μ_u, μ_d	0.5	0.9	0.4	1.2	1.5
3		ν	2.5	-5.0	1.4	-2.7	-2.2
4	Logistic	ν_u, ν_d	3.0	-4.1	1.7	-1.5	-0.7
5		α, β, γ	18.3	-32.5	3.9	-3.8	-15.5
6		$\alpha_u, \alpha_d, \beta, \gamma$	26.7	-47.4	4.0	-1.9	12.0
7		$\alpha_u, \beta_u, \beta_d, \gamma$	21.8	-37.6	4.3	-2.5	10.0
8		$\alpha_u, \alpha_d, \beta_u, \beta_d, \gamma$	29.0	-50.1 ^a	9.2	-10.3 ^a	14.8
9		$\alpha_u, \alpha_d, \beta_u, \beta_d, \gamma_u, \gamma_d$	29.0	-48.1	9.2	-8.3	14.8

Model	Parameters	DYS391 (N = 14595)		DYS392 (N = 14549)		DYS393 (N = 13178)	
		ΔML	ΔAIC	ΔML	ΔAIC	ΔML	ΔAIC
1	SMM	μ	0	0	0	0	0
2	Linear	μ_u, μ_d	0.4	0.1	1.9	0.1	1.7
3		ν	1.8	-3.7	0.3	-0.6 ^a	-1.3
4	Logistic	ν_u, ν_d	2.3	-2.5	0.4	1.2	0.4
5		α, β, γ	13.2	-22.4	1.0	2.1	-6.6 ^a
6		$\alpha_u, \alpha_d, \beta, \gamma$	20.9	-35.8	0.9	4.2	-6.1
7		$\alpha_u, \beta_u, \beta_d, \gamma$	19.4	-32.8	1.0	4.1	-6.1
8		$\alpha_u, \alpha_d, \beta_u, \beta_d, \gamma$	22.6	-37.3 ^a	3.2	1.7	-5.3
9		$\alpha_u, \alpha_d, \beta_u, \beta_d, \gamma_u, \gamma_d$	23.4	-36.9	3.2	3.7	-5.3

Differences in maximum log-likelihood (ML) and corresponding Akaike information criterion (AIC) values for some single-step microsatellite mutation models are shown. N: number of father-son duos with data available for the respective locus. The stepwise mutation model (SMM) with a single parameter was chosen as the reference model so that ΔML is the maximum log-likelihood minus this value for the symmetric SMM. Similarly, ΔAIC is the AIC value minus the value for the symmetric SMM.

^a For each locus, the smallest AIC value, indicating the best model.

Figure 2 Mutation probabilities of three models fitted to the DYS19 data. The top half of the diagram depicts upward mutation probabilities whereas the bottom half depicts downward mutation probabilities. Triangles mark the observed mutation frequencies (solid, upward; open, downward). Models fitted are the linear model without symmetry (red lines), the fully symmetrical logistic model (green lines), and the logistic model with parameters α_u , α_d , β_u , β_d , and γ (blue lines). Note that the plot is truncated beyond paternal allele 17 because the observed downward mutation frequency for paternal allele 18 is very high, but unreliable, due to the small number of fathers carrying allele 18. Mutation frequencies not shown are $\hat{p}_{18,19} = 0$, $\hat{p}_{19,20} = 0$, $\hat{p}_{18,17} = 0.15$, and $\hat{p}_{19,18} = 0$.

Discussion

It is well known that the pattern of microsatellite mutation varies across loci (Kelkar *et al.* 2008). To our knowledge, however, the present study is the first to systematically compare novel as well as previously described microsatellite mutation models for Y-STRs in a locus-specific fashion. This comparison was made possible by the accumulation of suitable genotype data from 15,285 father–son duos. The major advantage of this type of data is that all father–son relationships had been confirmed by independent genotyping of other markers. In studies using deep pedigree data for mutational analyses, this is typically not the case (Heyer *et al.* 1997), which renders discrimination between false paternity records and genuine mutations notoriously difficult. Furthermore, by basing our analysis on directly observed mutations (*i.e.*, Y-STR mutations in father–son duos), we avoided the need for additional assumptions about the underlying population dynamics, mating behavior, or selective pressure. This is a clear advantage over studies that sought to investigate microsatellite mutation processes by comparing distantly related genomes (Dieringer and Schlötterer 2003).

In our analysis, we also avoided the complicating effects of recombination through choosing loci from the male-

specific region of the Y chromosome. Although this restriction may at first glance seem to limit the general applicability of our results, it may be surmised that Y-chromosomal and autosomal microsatellite loci obey similar mutation models because they have similar biochemical properties and because replication slippage is responsible for STR mutations in both instances (Heyer *et al.* 1997; Kayser *et al.* 2000). This contrasts with minisatellite mutations, where recombination plays a significant role (Buard *et al.* 2000).

One caveat of our study is that the loci considered were originally selected for forensic applications because of their high variability. Therefore, we cannot exclude that our parameter estimates are biased toward higher mutation probabilities, but this seems unlikely to affect the general conclusion as to which models are most appropriate for microsatellites.

As was mentioned before, many statistical models have been proposed for the microsatellite mutation process (Calabrese and Sainudiin 2005). However, only a few of these turned out to be applicable to our data. For example, the model proposed by Kruglyak *et al.* (1998), which includes point mutations, was not deemed relevant to our study because, with the genotyping systems used in forensics, point mutations are not altering repeat counts (Gusmão *et al.* 2006). Moreover, we chose to restrict ourselves to one-step models owing to the scarcity of data on multistep mutations (Table 3). The three instances of mutations resulting in a change by 2 repeat units in our data were counted as single-step changes for model-fitting purposes. This concerned two of the six loci considered, for which the models should therefore be interpreted as dichotomizing all possible mutation events into up- and downward mutations, regardless of step size. However, since multistep mutations are very rare, this dichotomization should not affect our conclusions substantially. Nevertheless, should more data on multistep mutations become available in the future, a study of more sophisticated models may become worthwhile.

Our study was in part inspired by Whittaker *et al.* (2003), who were the first to suggest the use of maximum likelihood to fit microsatellite mutation models. However, as explained above, their exponential mutation model was not well defined, resulting in unbounded mutation probabilities. This problem does not occur with a logistic model, which to our knowledge has not been investigated before, because in the logistic model mutation probabilities are always bounded by parameter γ . Notably, for small repeat numbers, Whittaker's model and the logistic model are similar in that both entail an exponential increase in mutation probabilities with increasing repeat number. Qualitatively, the logistic model is also similar to the best models emerging from genome comparisons, *e.g.*, the PL1 model in Sainudiin *et al.* (2004). The main features in both instances are an allele-length-dependent mutation rate and a confinement to single-step mutations.

In principle, it would be possible to combine mutation models to obtain better fits. For example, a model with

linearly increasing upward mutation probabilities but with downward mutation probabilities according to the logistic model fits our data for DYS390 and DYS391 somewhat better than the logistic model alone ($\Delta AIC = -21.9$ and $\Delta AIC = -37.7$, respectively, cf. Table 5). However, we decided to focus on pure models here to not exacerbate the multiple-testing problem.

Practical applications of our results are vast because many uses of microsatellite data require estimates of the respective mutation probabilities. The logistic model, which was shown here to provide the best fit to empirical mutation data, is readily applicable to likelihood-based kinship analysis, phylogenetic analysis, and coalescence methods used in population genetics. Our statistical evaluation of mutation models may also contribute to a better understanding of the underlying biological mutation mechanisms. In particular, the fact that the combined models fit the data better than the original ones suggests possible differences between the mechanisms of upward and downward mutation.

With these applications in mind, gathering of further mutation data, for example, in an international STR mutation database, seems to be warranted. With a growing database, it will become possible to further refine parameter estimates as well as the models themselves.

Acknowledgments

We thank all authors whose Y-STR mutation data were included in the present study and Lutz Roewer and Sascha Willuweit for providing access to summary statistics for the Y Chromosome Haplotype Reference Database (Willuweit and Roewer 2007). We thank Raazesh Sainudiin and an anonymous reviewer for helpful suggestions.

Literature Cited

- Akaike, H., 1973 Information theory and an extension of the maximum likelihood principle, pp. 267–281 in *Second International Symposium on Information Theory*, edited by B. N. Petrov. Akademiai Kiado, Budapest.
- Ballard, D., C. Phillips, G. Wright, C. Thacker, C. Robson *et al.*, 2005 A study of mutation rates and the characterisation of intermediate, null and duplicated alleles for 13 Y chromosome STRs. *Forensic Sci. Int.* 155: 65–70.
- Berger, B., A. Lindinger, H. Niederstätter, P. Grubwieser, and W. Parson, 2005 Y-STR typing of an Austrian population sample using a 17-loci multiplex PCR assay. *Int. J. Legal Med.* 119: 241–246.
- Bianchi, N., C. Catanesi, G. Bailliet, V. Martinez-Marignac, C. Bravi *et al.*, 1998 Characterization of ancestral and derived Y-chromosome haplotypes of New World native populations. *Am. J. Hum. Genet.* 63: 1862–1871.
- Buard, J., A. Shone, and A. Jeffreys, 2000 Meiotic recombination and flanking marker exchange at the highly unstable human minisatellite CEB1 (D2S90). *Am. J. Hum. Genet.* 67: 333–344.
- Budowle, B., M. Adamowicz, X. Aranda, C. Barna, R. Chakraborty *et al.*, 2005 Twelve short tandem repeat loci Y chromosome haplotypes: genetic analysis on populations residing in North America. *Forensic Sci. Int.* 150: 1–15.
- Burnham, K. P., and D. R. Anderson, 2002 *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, New York.
- Calabrese, P., and R. Sainudiin, 2005 Models of Microsatellite Evolution. In: *Statistical Methods in Molecular Evolution*, edited by Nielsen, R., pp. 290–305. Springer-Verlag, New York.
- Calabrese, P. P., R. T. Durrett, and C. F. Aquadro, 2001 Dynamics of microsatellite divergence under stepwise mutation and proportional slippage/point mutation models. *Genetics* 159: 839–852.
- de Souza Góes, A., E. de Carvalho, I. Gomes, D. da Silva, E. Fonseca Gil *et al.*, 2005 Population and mutation analysis of 17 Y-STR loci from Rio de Janeiro (Brazil). *Int. J. Legal Med.* 119: 70–76.
- Decker, A., M. Kline, J. Redman, T. Reid, and J. Butler, 2008 Analysis of mutations in father-son pairs with 17 Y-STR loci. *Forensic Sci. Int. Genet.* 2: e31–e35.
- Dieringer, D., and C. Schlötterer, 2003 Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome Res.* 13: 2242–2251.
- Di Rienzo, A., A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin *et al.*, 1994 Mutational processes of simple-sequence repeat loci in human populations. *Proc. Natl. Acad. Sci. USA* 91: 3166–3170.
- Domingues, P., L. Gusmão, D. da Silva, A. Amorim, R. Pereira *et al.*, 2007 Sub-Saharan Africa descendants in Rio de Janeiro (Brazil): population and mutational data for 12 Y-STR loci. *Int. J. Legal Med.* 121: 238–241.
- Dupuy, B., R. Andreassen, A. Flønes, K. Tomassen, T. Egeland *et al.*, 2001 Y-chromosome variation in a Norwegian population sample. *Forensic Sci. Int.* 117: 163–173.
- Dupuy, B., M. Stenersen, T. Egeland, and B. Olaisen, 2004 Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. *Hum. Mutat.* 23: 117–124.
- Durrett, R., and S. Kruglyak, 1999 A new stochastic model of microsatellite evolution. *J. Appl. Probab.* 36: 621–631.
- Falush, D., and Y. Iwasa, 1999 Size-dependent mutability and microsatellite constraints. *Mol. Biol. Evol.* 16: 960–966.
- Feldman, M. W., A. Bergman, D. D. Pollock, and D. B. Goldstein, 1997 Microsatellite genetic distances with range constraints: analytic description and problems of estimation. *Genetics* 145: 207–216.
- Garza, J., M. Slatkin, and N. Freimer, 1995 Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Mol. Biol. Evol.* 12: 594–603.
- Ge, J., B. Budowle, X. Aranda, J. Planz, A. Eisenberg *et al.*, 2009 Mutation rates at Y chromosome short tandem repeats in Texas populations. *Forensic Sci. Int. Genet.* 3: 179–184.
- Goedbloed, M., M. Vermeulen, R. Fang, M. Lembring, A. Wollstein *et al.*, 2009 Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFISTR Yfiler PCR amplification kit. *Int. J. Legal Med.* 123: 471–482.
- Gusmão, L., P. Sánchez-Diz, F. Calafell, P. Martín, C. Alonso *et al.*, 2005 Mutation rates at Y chromosome specific microsatellites. *Hum. Mutat.* 26: 520–528.
- Gusmão, L., J. Butler, A. Carracedo, P. Gill, M. Kayser *et al.*, 2006 DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. *Forensic Sci. Int.* 157: 187–197.
- Heyer, E., J. Puymirat, P. Dieltjes, E. Bakker, and P. de Knijff, 1997 Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum. Mol. Genet.* 6: 799–803.
- Hohoff, C., K. Dewa, U. Sibbing, K. Hoppe, P. Forster *et al.*, 2007 Y-chromosomal microsatellite mutation rates in a popu-

- lation sample from Northwestern Germany. *Int. J. Legal Med.* 121: 359–363.
- Ihaka, R., and R. Gentleman, 1996 R: a language for data analysis and graphics. *J. Comput. Graph. Stat.* 5: 299–314.
- Jobling, M. A., and C. Tyler-Smith, 2003 The human Y chromosome: an evolutionary marker comes of age. *Nat. Rev. Genet.* 4: 598–612.
- Kashi, Y., and D. G. King, 2006 Simple sequence repeats as advantageous mutators in evolution. *Trends Genet.* 22: 253–259.
- Kayser, M., L. Roewer, M. Hedman, L. Henke, J. Henke *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.* 66: 1580–1588.
- Kelkar, Y. D., S. Tyekucheva, F. Chiaromonte, and K. D. Makova, 2008 The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res.* 18: 30–38.
- Kimmel, M., R. Chakraborty, D. N. Stivers, and R. Deka, 1996 Dynamics of repeat polymorphisms under a forward-backward mutation model: within- and between-population variability at microsatellite loci. *Genetics* 143: 549–555.
- Krawczak, M., and J. Schmidtke, 1998 *DNA Fingerprinting*. Springer-Verlag, New York.
- Kruglyak, S., R. T. Durrett, M. D. Schug, and C. F. Aquadro, 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.
- Kurihara, R., T. Yamamoto, R. Uchihi, S. Li, T. Yoshimoto *et al.*, 2004 Mutations in 14 Y-STR loci among Japanese father-son haplotypes. *Int. J. Legal Med.* 118: 125–131.
- Lai, Y., and F. Sun, 2003 The relationship between microsatellite slippage mutation rate and the number of repeat units. *Mol. Biol. Evol.* 20: 2123–2131.
- Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.
- Lee, H., M. Park, U. Chung, H. Lee, W. Yang *et al.*, 2007 Haplotypes and mutation analysis of 22 Y-chromosomal STRs in Korean father-son pairs. *Int. J. Legal Med.* 121: 128–135.
- Lessig, R., and J. Edelmann, 1998 Y chromosome polymorphisms and haplotypes in West Saxony (Germany). *Int. J. Legal Med.* 111: 215–218.
- Mayntz-Press, K. A., and J. Ballantyne, 2007 Performance characteristics of commercial Y-STR multiplex systems. *J. Forensic Sci.* 52: 1025–1034.
- Nauta, M. J., and F. J. Weissing, 1996 Constraints on allele size at microsatellite loci: implications for genetic differentiation. *Genetics* 143: 1021–1032.
- Nocedal, J., and S. J. Wright, 2006 *Numerical Optimization*. Springer-Verlag, New York.
- Ohta, T., and M. Kimura, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetics* 22: 201–204.
- Padilla-Gutiérrez, J. R., Y. Valle, A. Quintero-Ramos, G. Hernández, K. Rodarte *et al.*, 2008 Population data and mutation rate of nine Y-STRs in a mestizo Mexican population from Guadalajara, Jalisco, México. *Legal Med.* 10: 319–320.
- Pawitan, Y., 2001 *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, New York.
- Pestoni, C., M. Cal, M. Lareu, M. Rodríguez-Calvo, and A. Carracedo, 1999 Y chromosome STR haplotypes: genetic and sequencing data of the Galician population (NW Spain). *Int. J. Legal Med.* 112: 15–21.
- Pontes, M., L. Cainé, D. Abrantes, G. Lima, and M. Pinheiro, 2007 Allele frequencies and population data for 17 Y-STR loci (AmpFSTR Y-filer) in a Northern Portuguese population sample. *Forensic Sci. Int.* 170: 62–67.
- Roewer, L., 2009 Y chromosome STR typing in crime casework. *Forensic Sci. Med. Pathol.* 5: 77–84.
- Sainudiin, R., R. T. Durrett, C. F. Aquadro, and R. Nielsen, 2004 Microsatellite mutation models: insights from a comparison of humans and chimpanzees. *Genetics* 168: 383–395.
- Shriver, M. D., and R. A. Kittles, 2004 Genetic ancestry and the search for personalized genetic histories. *Nat. Rev. Genet.* 5: 611–618.
- Shriver, M. D., L. Jin, R. Chakraborty, and E. Boerwinkle, 1993 VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. *Genetics* 134: 983–993.
- Sánchez-Diz, P., C. Alves, E. Carvalho, M. Carvalho, R. Espinheira *et al.*, 2008 Population and segregation data on 17 Y-STRs: results of a GEP-ISFG collaborative study. *Int. J. Legal Med.* 122: 529–533.
- Soares-Vieira, J. A., A. E. Billerbeck, E. S. Iwamura, B. B. Mendonça, L. Gusmão *et al.*, 2008 Population and mutation analysis of Y-STR loci in a sample from the city of São Paulo (Brazil). *Genet. Mol. Biol.* 31: 651–656.
- Tsai, L., T. Yuen, H. Hsieh, M. Lin, C. Tzeng *et al.*, 2002 Haplotype frequencies of nine Y-chromosome STR loci in the Taiwanese Han population. *Int. J. Legal Med.* 116: 179–183.
- Turrina, S., R. Atzei, and D. De Leo, 2006 Y-chromosomal STR haplotypes in a Northeast Italian population sample using 17plex loci PCR assay. *Int. J. Legal Med.* 120: 56–59.
- Valdes, A. M., M. Slatkin, and N. B. Freimer, 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* 133: 737–749.
- Weir, B. S., A. D. Anderson, and A. B. Hepler, 2006 Genetic relatedness analysis: modern data and new challenges. *Nat. Rev. Genet.* 7: 771–780.
- Weissenbach, J., 1993 A second generation linkage map of the human genome based on highly informative microsatellite loci. *Gene* 135: 275–278.
- Whittaker, J. C., R. M. Harbord, N. Boxall, I. Mackay, G. Dawson *et al.*, 2003 Likelihood-based estimation of microsatellite mutation rates. *Genetics* 164: 781–787.
- Willuweit, S., and L. Roewer, 2007 Y chromosome haplotype reference database (YHRD): update. *Forensic Sci. Int. Genet.* 1: 83–87.
- Xu, X., M. Peng, Z. Fang, and X. Xu, 2000 The direction of microsatellite mutations is dependent upon allele length. *Nat. Genet.* 24: 396–399.

Communicating editor: M. A. Beaumont