BIOINFORMATICS

Vol. 26 ECCB 2010, pages i440–i445
doi:10.1093/bioinformatics/btq367

Maximum likelihood estimation of locus-specific mutation rates in Y-chromosome short tandem repeats

Osnat Ravid-Amir and Saharon Rosset*

Department of Statistics and Operation Research, Tel Aviv, Israel

ABSTRACT

Motivation: Y-chromosome short tandem repeats (Y-STRs) are widely used for population studies, forensic purposes and, potentially, the study of disease, therefore knowledge of their mutation rate is valuable. Here we show a novel method for estimation of site-specific Y-STR mutation rates from partial phylogenetic information, via the maximum likelihood framework.

Results: Given Y-STR data classified into haplogroups, we de-scribe the likelihood of observed data, and develop optimization strategies for deriving maximum likelihood estimates of mutation rates. We apply our method to Y-STR data from two recent papers. We show that our estimates are comparable, often more accurate than those obtained in familial studies, although our data sample is much smaller, and was not collected specifically for our study. Furthermore, we obtain mutation rate estimates for DYS388, DYS426, DYS457, three STRs for which there were no mutation rate measures until now.

Contact: saharon@post.tau.ac.il

1 INTRODUCTION

Microsatellites or Short Tandem Repeats (STRs) are repetitive stretches of DNA made of short sequence motifs (2–6 bp), repeating a variable number of times. STRs are very common in eukaryotic genomes, and are highly mutable, with changes in repeat count occurring with much higher mutation rates compared to other polymorphisms [as high as 2×10^{-3} in Y-chromosome STRs (Y-STR) as estimated by Heyer *et al*. (1997)], leading to allelic polymorphism. These properties have made them an efficient tool for identification in forensics and paternity tests (Hammer and Redd, 2006; Kayser and Sajantila, 2001) as well as for studying demographic history and population structure, especially Y-STRs (Contu *et al*., 2008). This is because most of the Y-chromosome does not undergo recombination, hence population polymorphism originates only from mutations, and individuals can be placed on a common phylogenetic tree, whose branches are marked by mutations. Besides their use in forensics and demographic studies, some autosomal microsatellites are also known to be involved in disease: expansion of specific microsatellites beyond a certain threshold has been known to cause diseases such as Fragile X syndrome, myotonic dystrophy and Huntington's disease (Ashley and Warren, 1995; Rubinsztein, 1999). Microsatellites are also known to hypermutate in some cancers (Thibodeau *et al*., 1993).

In order to understand microsatellites mutation mechanisms, reliable rate estimates for STRs, in general, and Y-STR, specifically, have long been considered of scientific interest. Different approaches have been utilized for the estimation of Y-STR mutation rates, some of which use direct counting such as counting mutation events in deep routed pedigrees with known history (Heyer *et al*., 1997), or in father–son pairs (Ge *et al*., 2009; Kayser *et al*., 2000), and also in sperm (Holtkemper *et al*., 2001). Other methods try to estimate the mutation rates indirectly; for example, Zhivotovsky *et al*., 2004, use the diversity of Y-STRs in modern-day population samples with documented population founding or population splitting events in the last 1000 years (Gypsies and Pacific Islanders). They use T_D, an estimator for a population divergence time based on inter- and intra-population variance of STR repeat number (Zhivotovsky, 2001), to estimate the average STR mutation rates. For the eight Y-STRs they consider they get an estimate of $6.9\times10^{-4}\pm5.9\times10^{-4}$ mutations per generation.

All of the different approaches observed significant variation in the mutation rates of different STRs. See for example the comparable mean and standard error for mutation rates as estimated by Zhivotovsky *et al*. (2004). However, there is a roughly 3-fold gap between the rates estimated from geneologies and those estimated from historical or phylogenetic data (Zhivotovsky *et al*., 2004, 2006). This has generated extensive attention in the literature, with some explanations offered by Zhivotovsky *et al*. (2006), but in our view it largely remains unresolved.

In this article, we propose a new approach for estimating Y-STR mutation rates. As opposed to previous approaches, which require extensive collection of data specifically for this purpose (father–son pairs, genealogies, populations with documented history, etc.), our approach takes advantage of data collected in population genetic studies. There is a long list of such studies which make use of Y-STR data (Contu *et al*., 2008; Hammer *et al*., 2009; Quintana-Murci *et al*., 2010), and they often also sequence a collection of unique event polymorphism markers, usually single nucleotide polymorphisms (SNPs), which place each sample in a well-defined region of the human Y-chromosome phylogenetic tree, referred to as a haplogroup (Hg). We show below how this partial knowledge of the phylogenetic relationship between samples affects the likelihood of the observed Y-STR lengths, and demonstrate how the resulting optimization problems can be solved to obtain maximum likelihood (ML) estimates of Y-STR mutation rates. In the simplest cases, the resulting ML estimation problem is generalized linear model (GLM), with a non-standard complementary log–log (CLL) link function. This has been previously demonstrated and applied on mitochondrial DNA (mtDNA) data by Rosset *et al*. (2008). However, as we show, in Y-STR data we can extract more detailed information from the data. We formulate and solve the resulting ML maximization as general convex optimization problems.

We apply our approach to a combined dataset of haplogroup-associated Y-STR data from two recent papers (Hammer *et al*., 2009; Quintana-Murci *et al*., 2010), comprising in total 3780 samples in 66 haplogroups. We then compare our estimates to the ones

*To whom correspondence should be addressed.

published in the Y-Chromosome Haplotype Reference Database, YHRD (Willuweit and Roewer, 2007). The YHRD estimates are based on simple counting of mutations in very large numbers of meioses on known geneologies (between 10 000 and 25 000 meioses for each Y-STR) and are based on genealogical data collected specifically for this purpose. We demonstrate that our estimates are comparable to those derived from genealogical data, but more reliable (as reflected by having tighter confidence intervals). We are also able to obtain estimates for Y-STRs not contained in YHRD. Our approach can easily be applied to much larger datasets by augmenting our current study with additional datasets collected from the population genetics literature, with no proactive data collection effort required.

2 METHODS

2.1 Y-STR data

Our data comes from two recent population genetics studies. The first (Hammer *et al*., 2009) is a study of the Y-chromosome landscape of Jewish Cohanim, compared to the general Jewish population and to gentiles within the Middle East and other regions. The authors genotyped 75 SNPs on the Y-chromosomes of 3674 individuals. This classified them into 64 unique Y-chromosome Hgs, according to the accepted nomenclature. They also genotyped 12 Y-STRs in all individuals: DYS19, DYS385a, DYS385b, DYS388, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS426 and DYS439. An additional two Y-STRs (DYS438, DYS457) were genotyped in most individuals.

The second study (Quintana-Murci *et al*., 2010) deals with a very different population: the Colored people of South Africa, who are an intriguing genetic mixture of African, European and Southeast Asian ancestries. In this article, the authors used SNPs to divide the 228 males sampled into 21 Hgs, and they typed 14 Y-STRs on 226 of them: DYS19, DYS385a, DYS385b, DYS388, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS426, DYS437, DYS438 and DYS439.

We use the combined dataset from both studies for our analyses, where we rely on the Hg inference by the authors of these two papers, and use the Y-STR data they generated as input to our method. We removed DYS385a and DYS385b from the analysis, since their values cannot be uniquely determined, and replaced DYS389II by DYS389B = DYS389II-DYS389I, as is typically done (Goff and Athey, 2006; Mohyuddin *et al*., 2001). Our pre-processing of the data further included merging the datasets, specifically combining data for the 10 Hgs that appeared in both datasets; removing Hgs which had less than two samples in the combined dataset; removing samples which had less than 11 Y-STRs observed.

After this process, we were left with data on 66 Hgs, with a varying number of observations on each Y-STR, as detailed in Table 1.

Table 1. Sample numbers and sources for various Y-STRs in our data

STR name	Samples	Source
DYS19, DYS388, DYS389I, DYS389B, DYS390, DYS391, DYS392, DYS393, DYS426, DYS439	3780	Hammer *et al*. (2009) and Quintana-Murci *et al*. (2010)
DYS437	226	Qunitana-Murci *et al*. (2010)
DYS438	2766	Both papers
DYS457	2834	Both papers

2.2 Statistical estimation approach

Denote by T the phylogenetic tree containing all 3780 samples. We do not have the full tree structure, but rather a haplogroup view of that tree, that is, the samples are grouped into haplogroups or paragroups, here referred to as Hgs. These Hgs represent terminal subtrees of the full tree T, but we are not given their internal structure (Fig. 1). For each sample we are given the number of subunit repeats of all, or part of the 13 STRs mentioned bellow. We assume that:

(1) The haplogroup classification of all sequences is known and correct. This implies that the Hgs represent disjoint subtrees, and the method for mapping samples to these Hgs is accurate. Both assumptions are in fact very reasonably for our data, since the Y chromosome tree is very confidently mapped, and the classification is based on multiple slowly mutating SNPs (The Y Chromosome Consortium, 2002).

(2) Mutation rates of the different STRs are independent. For our estimation methodology, we need not assume anything about the 'distribution' of the rates.

(3) Every STR i ($i=1,\ldots,13$) has a fixed Poisson rate, λ_i, with which the mutations occur. The λ_i is the same in all Hgs.

Let $t(T)$ be the total time of all branches of our phylogenetic tree T, then, according to our assumptions , the number of mutations on this tree in STR i in total time $t(T)$ is distributed Poisson($\lambda_i t(T)$). In the same manner let $T_1,\ldots,T_K$ represent the K Hg terminal subtrees of T, whose total time length of all the branches, $t_1,\ldots,t_k$ and inner structure are not known. Thus, m_{ik}, the number of mutations of STR i, in Hg k, is distributed Poisson($\lambda_i t_k$). If we had the internal structure of each subtree k (of Hg k), then we could directly count m_{ik}, and hence formulate the total log-likelihood of the data and estimate the parameters, using Poisson regression, through the usual ML framework. However, as aforementioned, we do not, so we do not observe the m_{ik}'s, but only observe the state (number of subunit repeats) of STR i in all samples (leaves) of Hg k.

Here we go about this problem using an extension of a method proposed by Rosset *et al*. (2008) for SNP rate estimation in mtDNA. Briefly, given all states of STR i, in the leaves of Hg k, if the STR state is not identical in all leaves, we know for certain that $m_{ik}>0$, i.e. STR i has mutated at least once somewhere on the phylogenetic tree describing haplgroup k samples. If all of the samples in Hg k have the same number of subunits in STR i, we can conclude with almost absolute confidence that this site has not mutated anywhere on the Hg's phylogenetic tree, i.e. $m_{ik}=0$. To demonstrate that our approach can properly capture whether a mutation did occur in a specific site, consider a simple phylogenetic tree like the one in Figure 2, where we assume a mutation from red triangle to black circle has occurred on the top-right branch. The shapes at the bottom describe the states of the leaves (observed samples), if no other mutations have occurred at this site. Now assume we want all the leaves of the tree to have the same number of subunits (all triangles or all circles) at this STR. This would clearly require that either the mutation reverted back from circle to triangle on a cut of the subtree below the original mutation (such as both branches marked with **) or the same exact mutation (triangle to circle) simultaneously happened on a set of branches completing a cut of the full tree (such as the branch marked with X). If none of these highly unlikely events (requiring multiple

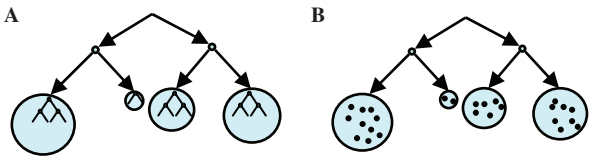

Fig. 1. (**A**) The full phylogenetic tree, including the internal Hg phylogenies, which we assume we do not observe. (**B**) Schematic of the Hg view of a phylogenetic tree.

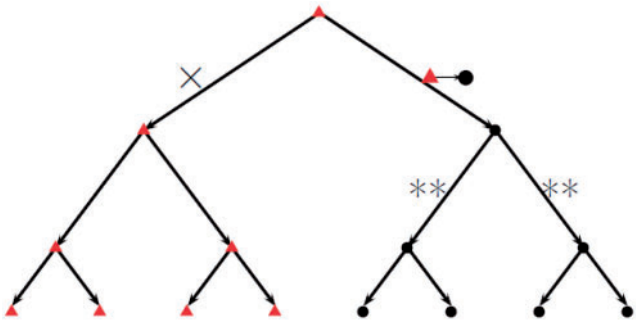

Fig. 2. Demonstration of our reasoning that we know whether any mutations have occurred in a specific STR. A mutation from red triangle to black circle has occurred on the top right branch. The shapes at the bottom describe the states of the leaves (observed samples). In order for such a mutation to go unnoticed all the leaves at the bottom should have the same state (all circle, or all triangle). Therefore, either both branches marked '**' changed back, or the same exact mutation happened also at the branch marked X. Both events are highly unlikely.

'coordinated' mutations) occur, all leaves would not have the same state at this site, given the shown triangle to circle mutation. We can illustrate the low probability of missing a mutation in our approach, by comparing it to another probability that of not observing a mutation on a coalescent tree because it has mutated back on the same link and thus is completely unobservable. Assuming for simplicity that all polymorphisms are binary, consider for example the two links marked with **, assume they both have length t. It is easily seen that the probability that STR i mutated and reverted on either one of them is $2\exp(-2\lambda_i t)(\lambda_i t)^2/2+O((\lambda_i t)^3)$. The probability that the triangle to circle mutation reverted back on both of them simultaneously is similarly $\exp(-2\lambda_i t)(\lambda_i t)^2/2+O((\lambda_i t)^4)$, i.e. slightly smaller. If we do not assume both links have the same length, then the first probability is potentially much bigger than the second. Therefore, under reasonable assumptions, that reversion back is most likely on binary splits, our total chance of setting $m_{ik}=0$ when the true value is $m_{ik}>0$, is on the same order of magnitude as twice the chance that the coalescent tree contains mutations that reverted back on the same link, which are inherently unobservable (Rosset *et al.*, 2008).

Thus, while we could not observe the Poisson mutation counts m_{ik}, we observe the binary variables:

$$b_{ik}=\begin{cases}1 \text{ If } m_{ik}=0\\ 0 \text{ If } m_{ik}>0\end{cases}.$$

These variables are distributed as $b_{ik}\sim$ Bernoulli($\exp(-\lambda_i t_k)$). After formulating the partial log-likelihood [Equation (1)] of the observed data b, the ML estimation of the parameters is a binomial regression with a CLL link function. This is therefore a GLM problem (MacCullagh and Nelder, 1991), and can be solved using the standard GLM framework (Rosset *et al.*, 2008). Note that we use observed data only on m_{ik}, whose distribution depends only on the total branch length of each Hg subtree rather than on a specific internal structure. Hence the rate estimates will not depend on the internal structure as long as the total branch length is the same. We refer to this estimation procedure as MAL1 (ML estimation with information on at least one mutation).

$$\ell(b,\lambda,t)=\sum_{i,k}\left[-\lambda_i t_k b_{ik}+\log(1-\exp(-\lambda_i t_k))(1-b_{ik})\right] \quad (1)$$

However, using this method we do not use all data available, specifically how many different STR states are present in each haplogroup. Using the same reasoning as above, if two different states indicate that at least one mutation occurred, three different states indicate the presence of at least two mutations, etc. (with the maximal number of STR states per Hg in our data being 9). Since we assume that the number of mutation events of STR i in *Hg* k, $m_{ik}\sim$ Poisson($\lambda_i t_k$), the probabilities in each case can therefore be written as:

$$\begin{aligned}&P(m_{ik}=0)=\exp(-\lambda_i t_k)\\&P(m_{ik}>0)=1-\exp(-\lambda_i t_k)\\&P(m_{ik}>1)=1-\exp(-\lambda_i t_k)-\lambda_i t_k\exp(-\lambda_i t_k)\\&\vdots\\&P(m_{ik}>n)=1-\sum_{j=0}^{n}(\lambda_i t_k)^j\exp(-\lambda_i t_k)/j!\end{aligned}$$

Let y_{ik} be the observed number of states of STR i in *Hg* k. We can formulate the log-likelihood of the data y:

$$\begin{aligned}\ell(y,\lambda,t)=\sum_{i,k}\Big[&(I_{[(y_{ik}-1)=0]}\log(P(m_{ik}=0))\\&+\sum_{j=1}^{8}I_{[(y_{ik}-1)=j]}\log(P(m_{ik}>j-1))\Big].\end{aligned} \quad (2)$$

We refer to this estimation procedure as MAL8 (ML estimation with information on at least eight mutations). This does not fit the GLM framework, but is a convex function, and therefore has a global maximum, which can be found using standard optimization tools. Here we used the Matlab function 'fmincon' to find the optimum solution.

MAL1 and MAL8 procedures yield ML estimates of both the Hg tree lengths t_k; $k=1,\ldots,K$, and the STR-specific mutation rates λ_i; $i=1,\ldots,I$. However, note that this ML solution is defined only up to a multiplication of all the $\lambda_i s$ by a constant and division of all the $t_k s$ by the same constant. Thus, to complete the estimation we need to resolve this remaining degree of freedom. Here we calibrate our rates to the YHRD, setting the sum of STR rates that are in common equal. The STRs that are missing in the YHRD were multiplied by the same constant that was used for the calibration of the others.

2.2.1 Saturation and sub-sampling It so happens that some of the Hgs in our data are saturated, that is, for some *Hg* k all the STRs have more than one state: $y_{ik}>1\forall i$. This happens especially for Hgs that contain many individuals. In this case t_k is not estimable in our methodology (that is, the ML estimate is not finite). In order to reduce the amount of saturation we created different datasets by sub-sampling 50% of the samples belonging only to the 'problematic' Hgs, multiple times ($n=10$). We then applied our methodology to the different datasets and generated a 'distribution' of estimates. Our rate estimates proved to be very robust and changed very little in the different sets of data sampled. For each rate parameter, λ_i, the mean of the distribution was taken to be our final estimate.

2.3 Statistical inference

2.3.1 Bias and variance estimation based non-parametric bootstrap In order to assess how reliable our mutation rate estimates are, we use the Non-Parametric bootstrap (Efron and Tibshirani, 1993). Namely, we resample our data (with replacement) over and over again, 750 times, and run our estimation procedure (subsampling 10 times to get a distribution of estimates and taking the mean as the estimate $\lambda_i{}^{*b}$) on each bootstrap sample to get a distribution of bootstrap estimates. The central assertion of the bootstrap method is that the relative frequency distribution of these $\lambda_i{}^{*b}$'s is an estimate of the sampling distribution of the true λ_i. Hence, we can use this distribution to estimate the bias and variance of our estimate, and consequently for constructing confidence intervals.

Table 2. MAL1 and MAL8 mutation rate estimates and bootstrap based estimates of the bias and SD of the rate estimates

STR	Rate estimates		Bias		SD	
	MAL1	MAL8	MAL1	MAL8	MAL1	MAL8
DYS.19	2.49E-03	3.03E-03	3.67E-04	**−1.29E-04**	5.09E-04	**4.90E-04**
DYS.388	8.27E-04	1.02E-03	1.95E-04	**−7.65E-05**	1.82E-04	**1.59E-04**
DYS.389a	1.96E-03	2.04E-03	4.40E-04	**2.75E-04**	3.59E-04	**3.22E-04**
DYS.389b	2.95E-03	3.18E-03	2.19E-04	**6.35E-05**	6.14E-04	**5.71E-04**
DYS.390	2.41E-03	2.80E-03	3.96E-05	−3.05E-04	3.61E-04	**3.59E-04**
DYS.391	1.50E-03	1.04E-03	−2.52E-04	**5.05E-05**	1.89E-04	**1.66E-04**
DYS.392	6.95E-04	7.96E-04	1.34E-04	**−1.12E-05**	1.16E-04	**1.11E-04**
DYS.393	1.42E-03	1.77E-03	3.11E-04	**−2.62E-05**	2.90E-04	**2.76E-04**
DYS.426	1.07E-04	7.52E-05	1.96E-05	**9.34E-06**	3.07E-05	3.77E-05
DYS.437	8.92E-04	1.38E-03	6.09E-04	**2.34E-04**	3.20E-04	**2.78E-04**
DYS.438	9.81E-04	1.16E-03	2.77E-04	**−4.60E-05**	2.37E-04	**2.31E-04**
DYS.439	5.38E-03	3.71E-03	−1.93E-03	**−4.21E-05**	6.84E-04	**6.20E-04**
DYS.457	7.16E-04	7.28E-04	2.89E-04	**8.15E-05**	2.43E-04	**2.29E-04**

Bold values indicate bias/SD estimates which are lower using MAL8.

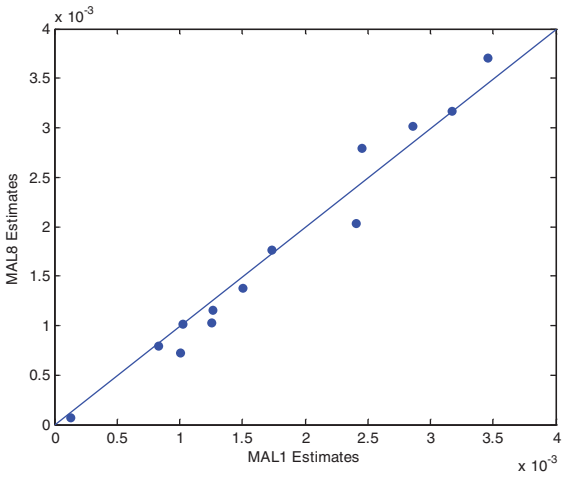

Fig. 3. Rate estimates of MAL1 compared to MAL8. Pearson correlation coefficient of 0.9162, Spearman 0.9560 (Hollander and Wolfe, 1973).

3 RESULTS

3.1 MAL1 and MAL8 estimated Y-STR mutation rates

In order to obtain estimation of Y-STR mutation rates, we used previously published datasets comprised of Y-STR lengths of 13 Y-STRs of >3780 individuals, assigned to one of 66 haplogroups (see Section 2). By using the length distribution of Y-STRs in each haplogroup, we apply the proposed MAL1 and MAL8 likelihood maximizations (see Section 2) to obtain uncalibrated mutation rate and branch length estimates. The estimations were then calibrated according to the YHRD.

Table 2 gives estimates of the 13 Y-STR mutation rates obtained using MAL1 and MAL8. Both methods give similar results, (Fig. 3), however, since the MAL8 method uses more information, we would expect it to give better results. This is indeed the case as both the SD and the bias estimated by the bootstrap are smaller for MAL8 compared to MAL1 in 12 out of the 13 STRs. Both the bias and SD give a significant P-value of 0.006 using a sign test (Wackerly *et al.*, 2002), indicating that MAL8 is indeed more accurate. Based on these observations we proceed using MAL8 for the estimation and inference of Y-STR rates.

Table 3. Rate estimates and confidence intervals (CI) of YHRD and MAL8

STR	YHRD	MAL8	YHRD		MAL8	
	Estimates	Estimates	CI		CI	
DYS.19	2.30E-03	3.03E-03	0.0016	0.0032	0.0022	0.0040
DYS.388	–	1.02E-03	–	–	0.0007	0.0013
DYS.389I	2.52E-03	2.04E-03	0.0017	0.0035	**0.0015**	**0.0027**
DYS.389B	3.64E-03	3.18E-03	0.0027	0.0048	0.0022	0.0044
DYS.390	2.10E-03	2.80E-03	0.0014	0.0030	**0.0021**	**0.0036**
DYS.391	2.60E-03	1.04E-03	0.0018	0.0036	**0.0008**	**0.0014**
DYS.392	4.12E-04	7.96E-04	0.0002	0.0009	**0.0006**	**0.0011**
DYS.393	1.05E-03	1.77E-03	0.0006	0.0018	**0.0013**	**0.0023**
DYS.426	–	7.52E-05	–	–	0.00003	0.0001
DYS.437	1.23E-03	1.38E-03	0.0006	0.0021	**0.0009**	**0.0020**
DYS.438	3.06E-04	1.16E-03	0.0001	0.0009	0.0007	0.0017
DYS.439	5.21E-03	3.71E-03	0.0039	0.0068	**0.0026**	**0.0050**
DYS.457	–	7.28E-04	–	–	0.0004	0.0012

CIs in bold indicate CIs which are tighter using MAL8.

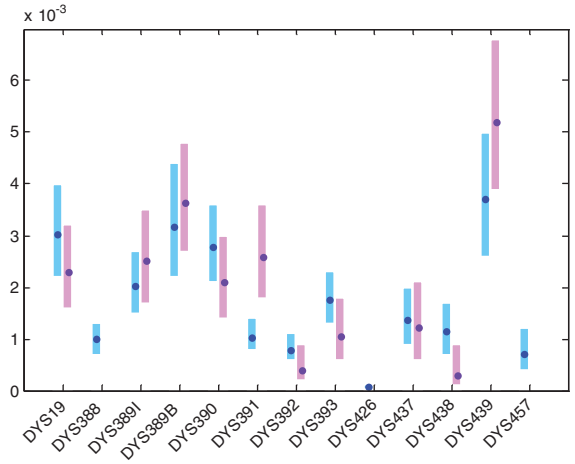

Fig. 4. Estimates and CIs of MAL8 (light blue) and YHRD (pink).

3.2 Inference and MAL8 estimate comparison with the YHRD estimates

Table 3 and Figure 4 show the rate estimates and confidence intervals of our method and the YHRD for each of the 13 STRs. Confidence intervals were calculated using 750 bootstrap samples, as described in the Section 2. While the rates vary considerably between the different STRs (ranging from 7×10^{-5} mutations per meiosis for DYS426 to about 3×10^{-3} for DYS19, 439 and 389B), the MAL8 predictions are similar to the YHRD measured rates, the only exception being DYS391. In seven out of 10 STRs, the confidence intervals obtained by MAL8 are a tighter than the YHRD confidence intervals.

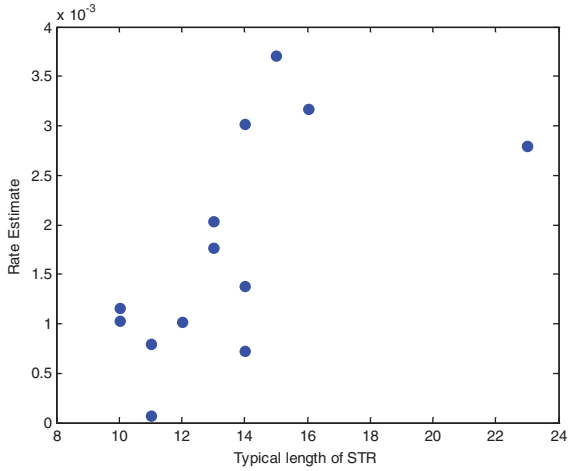

Fig. 5. Rate estimates of each STR plotted versus its typical length (mode).

In addition to the 10 STRs with known rates, we obtained predictions for three more Y-STRs: DYS388, DYS426, DYS457 with mutation rates of $1\times10^{-3}\pm1.59\times10^{-4}$, $7.5\times10^{-5}\pm3.7\times10^{-5}$, $7.3\times10^{-4}\pm2.29\times10^{-4}$ mutations per meiosis, respectively.

3.3 STR length and its rate

It has been previously suggested that STR mutation rate increases with the number of subunits (Calabrese and Sainudiin, 2005). Figure 5 shows each STR's mutation rate plotted against its modal length (the length appearing in the largest number of the samples). As can be seen there is a positive correlation, measured by Spearman's rho of 0.6713, with a significant *P*-value of 0.012. Thus, the estimated rates support this assertion.

4 DISCUSSION

In this study we show a new method for estimating Y-STR mutation rates. This method (MAL8) is based on ML estimation using partial phylogenetic information. Using data for 10 Y-STRs with previously estimated mutation rates, MAL8 obtains rates highly similar to the previously measured rates described in the YHRD (Willuweit and Roewer, 2007). This is noteworthy, as both the estimation method and the datasets used are very different.

Classical mutation rate estimation is based on counting mutations in father son pairs (or other known geneologies). This requires both a specialized dataset, and, due to the small rate of mutations, thousands of samples (YHRD uses 10 000–25 000 meioses for each STR). The MAL8 method used in this study was applied to a small, previously published dataset, consisting of only <4000 samples, which were collected for other uses and not customized for Y-STR mutation rate estimation in any way. Even on such a small dataset, the MAL8 method obtained tighter confidence intervals in most STRs, including DYS437 for which we had significantly fewer samples than the other Y-STRs (∼230 samples). This emphasizes an important aspect of our approach, whose performance largely depends on the level of detail in Hg classification, more than on the actual number of samples used. Thus, our approach is expected to do well in the presence of a detailed Hg phylogeny, even with relatively small sample sizes. Note that an incorrect classification into the Hgs might cause a contradiction to the assumption of independent occurrences of mutations, but this is not a real concern, since Hg classification is done with very high confidence according to SNPs (Hammer *et al*., 2009; Karafet *et al*., 2008; The Y Chromosome Consortium, 2002). Importantly, the rate estimates depend only on polymorphisms observed, which in turn depend on the total branch lengths of each Hg subtree, rather than on a specific internal structure.

We obtained mutation rate estimates for three STRs for which, to the best of our knowledge, there is no measured mutation rate: DYS388, DYS426 and DYS457. The estimated mutation rate of DYS426 (0.0007) is significantly slower than the other estimates. Indeed, we observed in this study, as well as others in previous works (Willuweit and Roewer, 2007), that there is the large difference in mutation rates between different STRs. This has been suggested to originate from differences in STR repeat counts (a connection which we have verified above exists in our data), the length of the subunit itself, or its nucleotide composition. Obtaining better statistics on the mutation rates in different STRs may enable to shed more light on this problem. While currently >400 Y-STRs have been described (Hanson and Ballantyne, 2006; Kayser *et al*., 2004), only <20 Y-STRs have measured rates (Willuweit and Roewer, 2007). Since the dataset used for the MAL8 is publicly available and continues to grow, obtaining sufficient data for additional STRs may enable estimation of their mutation rates. In addition, a current limiting factor for more accurate estimation is the resolution of the haplogroup classification. Studies with detailed Hg classification based on unique event polymorphisms like SNPs would be most useful for generating accurate estimates of Y-STR rates.

A fundamental flaw of all Y-STR mutation rate estimation approaches, including our own, is the limiting assumption they make about the nature of Y-STR mutation processes. The mutation count is assumed to be a symmetric random walk, so that the probability of change in the repeat count for each STR is fixed, independently of whether the change is an increase or decrease, and of the current repeat count. These models are unrealistic as they do not allow a stationary distribution of repeat counts (Calabrese and Sainudiin, 2005), thus, eventually leading to STR length which is infinite or zero—both possibilities are obviously unreasonable. Interestingly, it has been demonstrated that use of such simplistic models is consistent with a decrease in 'observed' mutation rates as distance between samples increases (Calabrese and Sainudiin, 2005). We believe that this should be further investigated as a possible factor in the 3-fold gap between the genealogical and evolutionary rates (Zhivotovsky *et al*., 2004).

Hence, we plan future extensions for our modeling approach to allow for asymmetric rates and length dependencies. Note that this is much more complex than our current approach, since knowledge of the total mutation count m_{ik} no longer suffices to write the likelihood. At the very least, this approach requires prior knowledge about ancestral STR length in every Hg, or, possibly, a way to infer it.

ACKNOWLEDGEMENTS

Special Thanks to Amnon Amir for his helpful ideas throughout this work.

Funding: European Union (grant MIRG-CT-2007-208019 to O.R.-A. and S.R., in part); Israeli Science Foundation (grant 1227/09 to O.R.-A. and S.R., in part).

Conflict of Interest: none declared.

REFERENCES

Ashley,C.T. and Warren,S.T. (1995) Trinucleotide repeat expansion and human disease. *Annu. Rev. Genet.*, **29**, 703–728.

Calabrese,P. and Sainudiin,R. (2005) Models of microsatellite evolution. In Nielsen,R. (ed.) *Statistical Methods in Molecular Evolution*. Springer, New York, pp. 290-305.

Contu,D. *et al*. (2008) Y-Chromosome based evidence for pre-neolithic origin of the genetically homogeneous but diverse sardinian population: inference for association scans. *PLoS ONE*, **3**, e1430.

Efron,B. and Tibshirani,R.J. (1993) *An Introduction to the Bootstrap*. Chapman and Hall, NY.

Ge,J. *et al*. (2009) Mutation rates at Y chromosome short tandem repeats in Texas populations. *Forensic Sci. Int. Genet.*, **3**, 179–184.

Goff,P. and Athey,T. (2006) Diagnostic Y-STR markers in haplogroup G. *J. Genet. Geneal.*, **2**, 12–17.

Hammer,M. and Redd,A.J. (2006) Forensic applications of Y chromosome STRs and SNPs. *Forensics in Law Enforcement*, **133**.

Hammer,M. *et al*. (2009) Extended Y chromosome haplotypes resolve multiple and unique lineages of the Jewish priesthood. *Hum. Genet.*, **126**, 707–717.

Hanson,E.K. and Ballantyne,J. (2006) Comprehensive annotated STR physical map of the human Y chromosome: Forensic implications. *Leg. Med.*, **8**, 110–120.

Heyer,E. *et al*. (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum. Mol.Genet.*, **6**, 799.

Hollander,M. and Wolfe,D. (1973) *Nonparametric Statistical Methods*. Wiley, New York.

Holtkemper,U. *et al*. (2001) Mutation rates at two human Y-chromosomal microsatellite loci using small pool PCR techniques. *Hum. Mol. Genet.*, **10**, 629–633.

Kayser,M. and Sajantila,A. (2001) Mutations at Y-STR loci: implications for paternity testing and forensic analysis. *Forensic Science International*, **118**, 116–121.

Kayser,M. *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.

Kayser,M. *et al*. (2004) A Comprehensive Survey of Human Y-Chromosomal Microsatellites. *Am. J. Hum. Genet.*, **74**, 1183–1197.

MacCullagh,P.J. and Nelder,J.A. (1989) *Generalized Linear Models*, 2nd edition, Chapman and Hall, London.

Mohyuddin,A. *et al*. (2001) Y-chromosomal STR haplotypes in Pakistani populations. *Forensic Sci. Int.*, **118**, 141–146.

Quintana-Murci,L. *et al*. (2010) Strong maternal khoisan contribution to the south african coloured population: a case of gender-biased admixture. *Am. J. Hum. Genet.*, **86**, 611–620.

Rosset,S. et al. (2008) Maximum-likelihood estimation of site-specific mutation rates in human mitochondrial DNA from partial phylogenetic classification. *Genetics*, **180**, 1511.

Rubinsztein,D.C. (1999) Trinucleotide expansion mutations cause diseases which do not conform to classical Mendelian expectations. In *Microsatellites: Evolution and Applications.* Oxford University Press, Oxford, England, pp. 80–97.

The Y Chromosome Consortium (2002) A nomenclature system for the tree of human y-chromosomal binary haplogroups. *Genome Res.*, **12**, 339–348.

Thibodeau,S.N. *et al*. (1993) Microsatellite instability in cancer of the proximal colon. *Science*, **260**, 816.

Wackerly,D.D. *et al*. (2002) Mathematical statistics with applications, 6th Edition, Pacific Grove, Duxbury, California.

Willuweit,S. and Roewer,L. (2007) Y chromosome haplotype reference database (YHRD): update. *Forensic Sci. Int. Genet.*, **1**, 83–87.

Zhivotovsky,L.A. (2001) Estimating divergence time with the use of microsatellite genetic distances: impacts of population growth and gene flow. *Mol. Biol. Evol.*, **18**, 700–709.

Zhivotovsky,L.A. *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.

Zhivotovsky,L.A. *et al*. (2006) Difference between evolutionarily effective and germ line mutation rate due to stochastically varying haplogroup size. *Mol. Biol. Evol.*, **23**, 2268.