Mutation rate estimates for 110 Y-chromosome STRs combining population and father-son pair data.

Concetta Burgarella, Miguel Navascués

Mutation rate estimates for 110 Y-chromosome STRs combining population and father-son pair data

Concetta Burgarella[1,2] & Miguel Navascués[1,3*]

[1] CNRS UMR 7625 Écologie et Évolution, École Normale Supérieure, 46 rue d'Ulm 75230 Paris (France).

[2] INIA, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria. Carretera de La Coruña km 7.5, 28040, Madrid (Spain).

[3] INRA, UMR CBGP (INRA/IRD/Cirad/Montpellier SupAgro), Campus International de Baillarguet, CS 30016, F-34988 Montferrier-sur-Lez Cedex, France

* Corresponding author: INRA, UMR CBGP (INRA/IRD/Cirad/Montpellier SupAgro), Campus International de Baillarguet, CS 30016, F-34988 Montferrier-sur-Lez Cedex, France

phone: +33(0)4.99.62.33.42

fax: +33(0)4.99.62.33.45

e-mail: navascues@supagro.inra.fr

Running title: Mutation rate estimates for Y-STRs

ABSTRACT

Y-chromosome microsatellites (Y-STRs) are typically used for kinship analysis and forensic identification as well as for inferences on population history and evolution. All applications would greatly benefit from reliable locus-specific mutation rates, to improve forensic probability calculations and interpretations of diversity data. However, estimates of mutation rate from father-son transmissions are available for few loci and have large confidence intervals, due to the small number of meiosis usually observed. By contrast, population data exist for many more Y-STRs, holding unused information about their mutation rates. To incorporate single locus diversity information into Y-STR mutation rate estimation, we performed a meta-analysis using pedigree data for 80 loci and individual haplotypes for 110 loci, from 29 and 93 published studies respectively. By means of logistic regression we found that relative genetic diversity, motif size and repeat structure explain the variance of observed rates of mutations from meiosis. This model allowed us to predict locus-specific mutation rates (mean predicted mutation rate 2.12×10^{-3}, SD = 1.58×10^{-3}), including estimates for 30 loci lacking meiosis observations and 41 with a previous estimate of zero. These estimates are more accurate than meiosis based estimates when a small number of meiosis is available. We argue that our methodological approach, by taking into account locus diversity, could be also adapted to estimate population or lineage specific mutation rates. Such adjusted estimates would represent valuable information for selecting the most reliable markers for a wide range of applications.

Keywords: mutation rate, Y-chromosome microsatellites, meiosis, population genetics, glm

INTRODUCTION

Around four hundred microsatellite markers from the Y human chromosome have been made available to date (e.g.[1]), with important applications in forensic analyses as well as in genealogy research. However, reliable locus-specific mutation rates are needed to carefully choose loci to minimize the error rate in kinship analysis and sample identification [2] while obtaining the maximum discriminatory power (e.g.[3-5]). Also in population genetics and evolutionary studies, correct inferences on the timing of major demographic events, the age of the most common ancestor, as well as dating Y-lineages and tracing disease evolution are based on the knowledge of mutation rates (e.g. [6-8]).

Population genetic theory predicts the genetic diversity of loci in function of their mutation rates (μ) and the effective size of populations (N). Therefore, it is possible to obtain estimates of the joint parameter $\theta=2N\mu$ from genetic diversity indices. In the case of loci evolving under a stepwise mutation model (SMM, generally assumed for microsatellites) it is possible to use the variance (V) in allele repeat count (i.e. allele size measured in number of repeats) and the 'homozygosity' ($H = \sum_{i=1}^{k} p_i^2$, where k is the number of different alleles in the population and p_i is the frequency of the i^{th} allele; note that the term homozygosity is not biologically meaningful for haploid loci but it will be used through the article for the sake of simplicity) for the estimation of θ using the following relationships [9]:

$$\hat{\theta}_V = \hat{V} \qquad [1]$$

67 $$\hat{\theta}_H = \frac{1}{2}\left(\frac{1}{\hat{H}^2} - 1\right) \quad [2]$$

68 where the hat denotes estimated values. However, because it is difficult to separate the
69 effects of demography (i.e. *N*), estimates of θ provide little information about mutation rate.
70 Nevertheless, it is possible to obtain information about relative mutation rates. In the case
71 of effective population size being the same among loci within one population (i.e. neutral
72 loci with same ploidy level, such as the Y-STRs), the ratio between the θ of two loci should
73 be the same as the ratio between their mutation rates [10]. However, relative mutation rate
74 estimates have limited utility for dating evolutionary events or calculating forensic
75 probabilities.
76
77 Absolute mutation rate estimates can be obtained by the analysis of allele transmissions in
78 pedigrees (e.g. [11,12]). The proportion of allele mismatches in father-son transmissions is
79 currently the most widely used approach to obtain estimates of mutation rates for Y-STRs.
80 Because of the low values of mutation rates, large number of father-son pairs must be
81 genotyped to obtain accurate estimates. This has limited the number of Y-STR loci for
82 which these estimates exist and many of them have been obtained from rather low sample
83 sizes. On the other hand, population diversity data exist for many more Y-STRs, holding
84 unused information about their mutation rates. The objective of this work is to present a
85 method to combine pedigree and population data for the estimation of mutation rates and to
86 provide locus-specific mutation rate estimates for 110 Y-STR loci (71 of which had no
87 previous estimate).
88

MATERIALS AND METHODS

Source of population data

Population data for 110 Y- chromosome microsatellite loci have been collected from 93 published works, for a total of 22,165 individual haplotypes (note that each individual was genotyped for a subset of loci and never for all of them). Locus names, sample sizes and references are detailed in Supplementary table S1. Locus nomenclature and allele call have been thoroughly checked to assure congruence across works and to remove duplicate data. Any population data with incongruent allele codes were either made uniform (when information provided by authors made it possible unequivocally) or excluded from analysis. Specifically, data from GATA H4 and GATA H4.1 have been pooled under the name GATA H4.1 by applying the appropriate correction to allele calls[13,14] and DYS389II has been transformed into DYS389B by subtracting allele size of DYS389I[15]. Multi-copy loci as well as single individuals with duplicated or variant alleles were excluded from the analysis. Data sets were chosen in order to obtain a maximum representation of loci and of geographical areas; collection of data stopped when no additional data sets could be found that would add data for new loci or would increase the order of magnitude of the sample size for individuals genotyped for a locus.

Source of meiosis (father-son pair) data

Direct observations of mutation events from meiosis data (father-son pairs) have been collected for 80 loci among the 110 loci with population data, from 29 published studies (table 1 and supplementary table S2). Confidence intervals from binomial probability distribution were estimated according to Wilson method[16]. Mutations assigned to DYS389II were carefully checked to discriminate those actually occurring in the DYS389I fragment from those occurred in the DYS389B fragment. Discrimination was always possible except for data from reference [11], which were excluded for this locus.

Statistical analysis

Population data were analyzed to obtain estimates of relative mutation rates between pairs of loci from allele repeat count variance and homozygosity. The relationship between relative mutation rates and meiosis based mutation rates was assessed by logistic regression using loci with both population and meiosis data. Inferred relationship was then used to predict mutation rates for all loci, including those lacking of meiosis data. Analysis procedure is detailed below.

First, we selected one locus to serve as reference (i.e. mutation rates for all other loci will be relative to this one). As mentioned above, not all individuals are genotyped for the same set of loci (cf. Supplementary table S1), thus it is not possible to use the whole data set in the logistic regression (although data from unused loci will be useful for predictions, see below). As a consequence, a reference locus has to be chosen in a way to maximize the amount of information used (i.e. to maximize the number of loci with meiosis participating

in the regression analysis). In other words, the reference locus has to be the one which shares genotype data with the greatest number of loci with meiosis. To achieve this, we used the following criteria (in this order): (i) there should be meiosis data for the reference locus, (ii) the number of loci with meiosis data (for at least 100 transmissions) and genotyped in individuals (from the population data) also genotyped for the reference locus should be maximum, (iii) the number of loci genotyped in individuals also genotyped for the reference locus should be maximum and (iv) the sampling size (number of individuals from population data) of the reference locus should be maximum. Note that the choice of the reference locus influences only the amount of data used in the analysis. Otherwise, the reference locus only sets an arbitrary scale to the relative mutation rates calculated from genetic diversity indices.

Relative mutation rate ($R = \mu_l / \mu_r$) for each locus *l* was estimated exclusively from individuals genotyped for both locus *l* and reference locus *r*. This ensures that the genetic diversity of the sample of both loci has been influenced by the same demographic history (this allows assuming the same effective population sizes). Thus, in the absence of selection, the differences in genetic diversity can be attributed solely to the mutation process. Moreover, because of the complete linkage of Y-STRs, data for both loci will share also the same exact genealogy (even if a selective process was in action). Because both loci have the same genealogy, estimates of the mutation rate ratios will be more accurate than in unlinked loci whose genealogies would vary largely due to the randomness of the coalescent process (e.g. nuclear STRs compared in reference[17]). Estimates $\hat{\theta}_{V,l}$, $\hat{\theta}_{V,r}$,

156 $\hat{\theta}_{H,l}$ and $\hat{\theta}_{H,r}$ were obtained from repeat count variance and homozygosity for loci *l* and *r*
157 (using equations 1 and 2) and two estimates of the relative mutation rate were calculated
158 from ratios $\hat{R}_V = \hat{\theta}_{V,l} / \hat{\theta}_{V,r}$ and $\hat{R}_H = \hat{\theta}_{H,l} / \hat{\theta}_{H,r}$.
159
160 A number of loci (24 out of 110, see supplementary table S1) for which there is population
161 data available, were not genotyped at the reference locus in any of the samples. For those
162 loci, relative mutation rates were estimated as described above but using the total number
163 of individuals available for each locus (we will denote these estimates $\hat{R}'_V$ and $\hat{R}'_H$). It
164 must be noted that $\hat{R}'_V$ and $\hat{R}'_H$ might have a larger error than $\hat{R}_V$ and $\hat{R}_H$ because the
165 effects of demography are more loosely accounted for. For this reason they were not used
166 for the estimation of the logistic regression model but only in the prediction of mutation
167 rates (see details below).
168
169 A generalized linear model (binary logistic regression[18]) was applied to the proportion of
170 mutations per meiosis. We tested for the relationship between meiosis mutation rate and
171 population relative mutation rates ($\hat{R}_V$, $\hat{R}_H$). Besides, some studies have proposed that
172 microsatellite mutation rates depend on allele length[19,20], motif size and motif structure[19].
173 Thus, in addition to $\hat{R}_V$ and $\hat{R}_H$, mean allele repeat count (*A*; estimated from the
174 population data), CG content in motif (P_{CG}; proportion of CG base pairs in the motif), and
175 the categorical variables motif size (*M*; tri-, tetra-, penta- or hexanucleotide motif) and

176 repeat structure (*S*; simple *versus* complex) were considered explanatory variables.
177 Information about Y-STR motifs was obtained from[21-24].
178
179 Problems of multicollinearity were evaluated on the full model (containing all explanatory
180 variables), as collinear variables represent partial redundant information and correlations
181 between variables generate unreliable individual estimates of regression coefficients.
182 Alternative models obtained after removing different combinations of collinear variables
183 were considered and reduced by stepwise removal of variables to minimize Akaike
184 information criterion (AIC, i.e. a standard procedure to find the explanatory variable
185 combination which accounts for the maximum of the variability with the minimum number
186 of variables). Reduced models were hereafter compared through their pseudo-R^2 value
187 (calculated by the maximum likelihood method[25]). Pseudo-R^2 measures the amount of
188 variation in the observed mutation rates explained by the model. The reduced model with
189 the highest pseudo-R^2 was chosen to predict mutation rates for all loci. As explained before,
190 for loci whose $\hat{R}_V$ and $\hat{R}_H$ could not be calculated, $\hat{R}'_V$ or $\hat{R}'_H$ were used as a proxy
191 (estimates for those loci will be distinguished in the results, as they are theoretically less
192 reliable).
193
194 All statistical analyses were performed in R[26], using packages *binom*[27] for calculation of
195 confidence intervals (CI), *ape*[28] for calculation of heterozygosity, and *pscl*[29] for calculation
196 of pseudo-R^2. A script in R language with the detailed analysis is available from the authors
197 upon request.

198

199 *Validation of the approach*

200

201 Performance of the statistical approach proposed was evaluated by means of simulations. In
202 each simulation a set of 108 fully linked loci were considered. Loci were divided in three
203 motif size categories: 36 'tri', 36 'penta' and 36 'tetra'. 'Tri' loci evolved at six different
204 mutation rates (10^{-4}, 2×10^{-4}, 4×10^{-4}, 8×10^{-4}, 1.6×10^{-3} and 3.2×10^{-3}, measured in mutations
205 per generation). 'Penta' loci evolved at mutation rates double to those for 'tri' loci (i.e.
206 2×10^{-4}, 4×10^{-4}, 8×10^{-4}, 1.6×10^{-3}, 3.2×10^{-3} and 6.4×10^{-3}) and 'tetra' loci evolved at mutation
207 rates quadruple to those for 'tri' loci (i.e. 4×10^{-4}, 8×10^{-4}, 1.6×10^{-3}, 3.2×10^{-3}, 6.4×10^{-3} and
208 1.28×10^{-2}). Note that categories 'tri', 'penta' and 'tetra' are arbitrary (both in their name
209 and their influence in mutation rate) and are only used to include the effect of a categorical
210 variable in the evaluation of the proposed approach. For each mutation rate within each
211 locus category, six loci differing in the amount of observed meiosis (i.e. 50, 150, 500, 1500,
212 5000 and 15000 meiosis) have been considered. To sum up, tree categories times six
213 mutation rates, times six loci differing in the number of meiosis gives 108 total simulated
214 loci.

215

216 Meiosis were simulated using the binomial distribution, with the probability equal to the
217 true mutation rate and the number of observations to the number of meiosis. Population
218 data were simulated with the coalescent simulator SimCoal2 [30] under a stepwise mutation
219 model. A sample size of 500 haplotypes was taken from a single population of constant
220 effective size of 1500 individuals (this effective size combined with the simulated mutation

rates yielded genetic diversity levels similar to those found on Y-STRs, i.e. around 2-14 alleles per locus).

Mutation rates estimates were obtained for each locus either by using exclusively meiosis data or by using a logistic regression on the observed mutations in meiosis using $\hat{R}_H$ and the simulated categorical variable ('motif size') as explanatory variables, according to the final model chosen with the real data (see results). The process was repeated 10000 times.

Root of the relative mean squared error ($RrelMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{\mu}_i - \mu)^2}{\mu^2}}$, where n is the number of simulations, $\hat{\mu}_i$ is the estimated mutation rate in simulation i and μ is the true mutation rate) was calculated for the two types of mutation rate estimates at each of the 108 loci.

RESULTS

Locus DYS643 was selected as reference locus following the criteria described above. Mutation rates relative to reference locus were estimated from repeat count variance and homozygosity for 86 loci, which were used in the logistic regression model (Table 1). Problems of multicollinearity were found between $\hat{R}_V$ and the mean repeat count (A), between $\hat{R}_H$ and A and between $\hat{R}_H$ CG content in motif (P_{CG}). Thus, we considered three alternative models with a different combinations of non-collinear variables each: model m1 including $\hat{R}_H$ plus the motif size (M) and repeat structure (S); model m2 including $\hat{R}_V$ plus

242 *M*, content in motif (P_{CG}) and *S*; m3 including *A* plus *M*, P_{CG} and *S*. The AIC minimization
243 approach led to the removal of variable P_{CG} in m2 and m3. Final models (supplementary
244 tables S3, S4 and S5) were ranked by their pseudo-R^2 values: 0.84 for reduced m1, 0.83 for
245 reduced m2, and 0.67 for reduced m3. Reduced m1 model was therefore selected to make
246 predictions on mutation rates from population data for all loci, using $\hat{R}_H$ or $\hat{R}'_H$. Reduced
247 m1 ($L\mu = \beta_0 + \beta_1 \hat{R}_H + \beta_2 M_{tri} + \beta_3 M_{tetra} + \beta_4 M_{penta} + \beta_5 S_{simple}$ + error; table S3 and figure 1)
248 shows that mutation rate estimated from meiosis ($L\mu$) increases with $\hat{R}_H$ (i.e. $\beta_1>0$),
249 depends on repeat size (highest for tetranucleotide loci followed by penta- and tri-, i.e. $\beta_3>$
250 $\beta_4> \beta_2$), and on the complexity of the loci (higher for simple than for complex loci, i.e.
251 $\beta_5>0$). Note that the coefficient of categorical variables is a value relative to the coefficient
252 of the category no explicitly represented in the equation (i.e. hexanucleotide repeat motif
253 class and the complex structure class).
254 For comparison, results from simple models (i.e. including each explanatory variable
255 separately) are reported in supplementary table S6. They show that all explanatory
256 variables, but the repeat structure, explain significantly part of the variability of mutation
257 rate estimates, although during the model minimization process some were excluded
258 because they provide redundant or non-independent information. Although repeat structure
259 is not able to significantly explain mutation rate variability when it is the only explanatory
260 variable, it is found to provide significant information when analyzed in combination with
261 other explanatory variables (supplementary tables S3, S4 and S5).
262

263 Predicted values for mutation rates range from 3.60×10^{-4} mutations per generation for
264 DYS645 to 9.64×10^{-3} for DYS449 (average 2.12×10^{-3}, SD = 1.58×10^{-3}; table 1). For those
265 loci which are not genotyped in any individual genotyped for the reference locus in the
266 population data (see table 1), differences in population history and genealogies are expected
267 to make an additional contribution to the variance in mutation rate estimates, although this
268 does not seem to be too important (exclusion of those loci hardly changes the average
269 predicted mutation rate, to 2.25×10^{-3}, SD = 1.65×10^{-3}). In total, regression approach
270 provides an estimate for 71 loci with either zero observed mutations in meiosis (i.e. point
271 estimate of mutation rate was zero) or lacking meiosis observations.
272
273 It is worth to notice that 45 out of 80 loci with meiosis data share their meiosis mutation
274 rate estimates and CI with at least another locus (given that often the same number of
275 mutations are observed in the same number of meiosis), while mutation rates predicted by
276 regression are different from each other for all loci. Simulations showed that the error
277 associated to meiosis mutation rate estimates is strongly influenced by the number of
278 meiosis, while the error of regression estimates seems independent of the number of
279 observed meiosis (figure 2 reports results for the four simulated mutation rates shared by all
280 loci categories, see methods). Error in both estimates depends on the true mutation rate,
281 decreasing for higher true mutation rates. However, this decrease is stronger for regression
282 estimates than for meiosis estimates. An interesting feature is that regression estimates are
283 more accurate than meiosis estimates when a low number of meiosis is available, but the
284 contrary occurs for high number of meiosis observations. Although this general pattern
285 seems to be independent of the true mutation rate, the threshold from which meiosis

286 estimates are more accurate than regression estimates increases with the true mutation rate.
287 It is important to remember that the behaviour described by simulations regards only loci
288 for which a meiosis estimate is available, however the regression approach provides an
289 estimate even when meiosis data are not available.
290
291 DISCUSSION
292
293 Mutation rates are expected to vary substantially across Y- microsatellite loci (reference [31]
294 and references therein). Such large variation has been attributed to motif size, complexity
295 of repeat structure and allele size (e.g.[12,21,32-34]). Our results are in general agreement with
296 the aforementioned works. We found that meiosis mutation rates are positively correlated
297 with population diversity (estimated by either homozygosity or relative repeat count
298 variance, tables S3 and S4), and mean repeat count (table S5), and depends on repeat motif
299 and repeat structure (table S3). The model selection approach used in this work indicates
300 that a model including relative genetic diversity (from homozygosity), repeat motif and
301 repeat structure as predictive variables is the best one to explain the variability found in
302 meiosis based estimates of mutation rate. However, it should be noted that the alternative
303 models tested (table S4 and S5) are valid too, although their lower pseudo-R2 values
304 indicate that they might have a lower performance for making inferences.
305
306 Correlation between mutation rates from meiosis and the relative mutation rates based on
307 homozygosity was positive and highly significant (tables S3 and S6). The latter is estimated
308 from population data, thus corresponds to the "evolutionary" mutation rates (i.e. the

309 effective mutation rate integrated over the history or gene tree of the sample). Pedigree
310 based mutation rate estimates have been shown to be up to 10-fold higher than evolutionary
311 mutation rate estimates (for sequence data), not only in Y-chromosomes (e.g.[31,35]) but also
312 in mitochondrial loci (e.g.[36,37]). The reasons for this discrepancy are still under discussion
313 and are likely to be found in the different temporal scale of estimation. In fact, slowly
314 mutating loci or reverse mutations as well as demographic fluctuations or differential
315 selection over generations are expected to affect population based diversity (see discussion
316 in reference [31]). It must be noted that our reported estimates (predicted from the logistic
317 model) correspond to the point estimates of mutation rates (i.e. mutation occurrence in
318 single generation).
319
320 Tri-, tetra- and pentanucleotide classes are well represented in the analyzed locus set (with
321 17, 55 and 7 loci respectively), while hexanucleotide class did not contribute much to the
322 regression model because it is present with only one locus (DYS448) with meiosis
323 observations. We found that the value for the model coefficient (β) is much lower for tri-
324 and pentanucleotide loci than for tetranucleotide loci (tables S3 and S4), which corresponds
325 to general lower mutation rates for tri- and pentanucleotide loci than for tetranucleotide loci
326 (table 1). Such a different behaviour is congruent with the results of previous studies. Järve
327 et al.[22] recently showed that pentanucleotide markers have two times lower repeat variance
328 and diversity than tetranucleotide markers, a feature probably related to a lower occurrence
329 of replication slippage with longer repeats. Regarding trinucleotide markers, Kayser et al.[21]
330 found they had often lower variance than tetranucleotide markers, probably because of the
331 effect of low absolute repeat allele lengths included in their sample[21]. Lower mutability of

shorter alleles compared with longer ones has been observed several times[32,33,12,34]. Accordingly, our results show that the variation in meiosis mutation rates could be significantly explained by mean repeat count (tables S5 and S6). Furthermore, when no diversity variable is included in the model, both repeat count and repeat motif contribute independently to explain the mutation rate variability (i.e. model m3, table S5).

The repeat structure explains very little the observed STR mutation rates (table S3), but it is maintained after model reduction using the AIC. However, the coefficient β for simple loci is positive in the final reduced m1 (table S3), while it is negative when the repeat structure is used as the only explanatory variable (table S6). Thus, the effect of the repeat structure on mutation rate is of difficult interpretation. Previous studies have failed to find a relationship of simple versus complex repeats with genetic diversity among loci[38,32,22]. These results might be due to the lack of effect of repeat structure. However, our qualitative classification of loci as 'simple' or 'complex' could be missing essential information of complex loci (i.e. differential length of the homogeneous array or combination of variable and constant repeats [21]) affecting the mutation rate. More precise definitions of the degree of complexity, similar to those used in ref. [21], could yield different results, but require detailed information on loci not readily available.

The model considered in this work for microsatellite evolution (SMM) predicts single-repeat-unit mutational changes. However, violations of this assumption have been reported both in phylogenetic and meiosis studies (e.g.[33,34,38,39]), suggesting that more complex models than SMM would better explain microsatellite variation. The ratio of variances in

number of repeats between two loci can be still considered a good estimator of the ratio of mutation rates even in case of multi-step mutations, provided that deviation from the SMM is similar for both loci (cf. equation 2 in reference[17]). Although the same argument is not strictly valid for the estimate of θ from homozygosity, small deviations from the SMM change very little the expected homozygosity (cf. Table I from [9]). Only 14 mutations (3.1% of total) involving multiple repeat units are included in meiosis data; therefore, SMM can be considered a reasonable approximation. In addition, the great congruence in prediction between models m1 (using homozygosity) and m2 (using repeat count variance) suggests that the mutation model violation is not an issue for the analysis (results not shown).

Some important outcomes derive from the approach proposed, emphasizing the positive impact of including population polymorphism data for the improvement of mutation rate estimates and the identification of loci distinctiveness. First, mutation rate estimates were obtained for 30 loci lacking estimates from meiosis observations. Second, locus-specific values of mutation rates can be obtained, while meiosis-based estimates give often equal values for several loci. Third, estimates can be obtained also for loci with very low mutation rates, for which a large sample of meiosis data is required to obtain a non-zero mutation rate estimate. Regarding this point, this work provides mutation rate estimates for 41 loci whose mutation rate estimate from meiosis was zero because no mutations had been observed. Lastly, regression based estimates present lower error than estimates from meiosis when only a 'low' number of meiosis is observed.

The analysis performed in this work represents a valuable tool for selecting most reliable markers to increase Y-STR set currently applied in forensic and kinship analyses. Also, the choice of adequate mutation rates keeps being an issue of great concern when inferences on human diversity and population history are pursued, as put in evidence in a recent work[6]. To account for the different variability of microsatellite loci, these authors use repeat count variance to obtain recalibrated evolutionary mutation rates for groups of loci. Our approximation allows more detailed results by achieving an adjusted mutation rate for each locus separately. The same methodology could be used to estimate population or lineage specific mutation rates, since different lineages and populations are often characterized by specific allele combinations [32,33,39] and mutation rate seems to be affected by allele size and structure. Finally, the analysis presented here can be easily automated for a data base, allowing the updating of estimates when new population and meiosis data are incorporated from upcoming studies.

AKNOWLEDGMENTS

We wish to thank Frantz Depaulis for his support and suggestions and Joaquín Navascués for its critical review of the manuscript. Two anonymous reviewers have contributed to improve the manuscript. This work was developed during a postdoctoral stay of CB (Bourse pour chercheurs étrangers de la Marie de Paris 2007).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

Supplementary information is available at European Journal of Human Genetics' website http://www.nature.com/ejhg/index.html

REFERENCES

1. Hanson EK, Ballantyne J. Comprehensive annotated STR physical map of the human Y chromosome: Forensic implications. *Leg. Med.* 2006;8(2):110–120.

2. Kayser M, Sajantila A. Mutations at Y-STR loci: implications for paternity testing and forensic analysis. *Forensic Sci. Int.* 2001;118(2–3):116–121.

3. Mulero JJ, Chang CW, Calandro LM, et al. Development and Validation of the AmpFℓSTR® Yfiler™ PCR Amplification Kit: A Male Specific, Single Amplification 17 Y-STR Multiplex System. *J. For. Sci.* 2006;51(1):64–75.

4. Lim S, Xue Y, Parkin E, Tyler-Smith C. Variation of 52 new Y-STR loci in the Y Chromosome Consortium worldwide panel of 76 diverse individuals. *Int. J. Legal Med.* 2007;121(2):124–127.

5. Vermeulen M, Wollstein A, van der Gaag K, et al. Improving global and regional resolution of male lineage differentiation by simple single-copy Y-chromosomal short tandem repeat polymorphisms. *Forensic Sci. Int.: Genet.* 2009;3(4):205–213.

6. Shi W, Ayub Q, Vermeulen M, et al. A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations. *Mol Biol Evol*. 2009:msp243.

7. Zerjal T, Wells RS, Yuldasheva N, Ruzibakiev R, Tyler-Smith C. A Genetic Landscape Reshaped by Recent Events: Y-Chromosomal Insights into Central Asia. *Am. J. Hum. Genet.* 2002;71(3):466-482.

8. Xue Y, Zerjal T, Bao W, et al. Male Demography in East Asia: A North-South Contrast in Human Population Expansion Times. *Genetics*. 2006;172(4):2431-2439.

9. Kimmel M, Chakraborty R. Measures of variation at DNA repeat loci under a general stepwise mutation model. *Theor. Popul. Biol.* 1996;50(3):345–367.

10. Xu H, Fu Y. Estimating effective population size or mutation rate with microsatellites. *Genetics*. 2004;166(1):555–563.

11. Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum. Mol. Genet.* 1997;6(5):799–803.

12. Gusmão L, Sánchez-Diz P, Calafell F, et al. Mutation rates at Y chromosome specific microsatellites. *Hum. Mutat.* 2005;26(6):520–528.

13. Gusmão L, Alves C, Beleza S, Amorim A. Forensic evaluation and population data on the new Y-STRs DYS434, DYS437, DYS438, DYS439 and GATA A10. *Int. J. Legal Med.* 2002;116(3):139–147.

14. Mulero JJ, Budowle B, Butler JM, Gusmão L. Nomenclature and allele repeat structure update for the Y-STR locus GATA H4. *J. For. Sci.* 2006;51(3):694–694.

15. Butler JM, Reeder DJ. Short Tandem Repeat DNA Internet DataBase. 2009. Available at: http://www.cstl.nist.gov/strbase/.

16. Wilson EB. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.* 1927;22(158):209-212.

17. Chakraborty R, Kimmel M, Stivers D, Davison L, Deka R. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proceedings of the National Academy of Sciences of the United States of America*. 1997;94(3):1041-1046.

18. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. 2nd ed. New York: Chichester, Wiley.; 2000.

19. Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am. J. Hum. Genet.* 1998;62(6):1408–1415.

20. Ellegren H. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat. Genet.* 2000;24(4):400–402.

21. Kayser M, Kittler R, Erler A, et al. A Comprehensive Survey of Human Y-Chromosomal Microsatellites. *Am. J. Hum. Genet.* 2004;74(6):1183-1197.

22. Järve M, Zhivotovsky LA, Rootsi S, et al. Decreased Rate of Evolution in Y Chromosome STR Loci of Increased Size of the Repeat Unit. *PLoS ONE*. 2009;4(9):e7276.

23. Gusmão L, Butler JM, Carracedo A, et al. 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.* 2006;157(2–3):187–197.

24. Leat N, Ehrenreich L, Benjeddou M, Cloete K, Davison S. Properties of novel and widely studied Y-STR loci in three South African populations. *Forensic Sci. Int.* 2007;168(2–3):154–161.

25. Long JS. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, California, USA: Sage Publications, Inc; 1997:297.

26. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2009. Available at: http://www.R-project.org.

27. Dorai-Raj S. binom: binomial confidence intervals for several parameterizations (version 1.0-5). 2009.

28. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*. 2004;20(2):289–290.

29. Jackman S. *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University*. Stanford, California: Department of Political Science, Stanford University; 2008. Available at: http://pscl.stanford.edu/.

30. Laval G, Excoffier L. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics*. 2004;20(15):2485–2487.

31. Zhivotovsky LA, Underhill PA, Cinnio lu C, et al. The Effective Mutation Rate at Y Chromosome Short Tandem Repeats, with Application to Human Population-Divergence Time. *Am J Hum Genet*. 2004;74(1):50–61.

32. Carvalho-Silva DR, Santos FR, Hutz MH, Salzano FM, Pena SD. Divergent Human Y-Chromosome Microsatellite Evolution Rates. *J. Mol. Evol.* 1999;49(2):204-214.

33. Dupuy BM, Stenersen M, Egeland T, Olaisen B. Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. *Hum. Mutat.* 2004;23(2):117–124.

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

35. Forster P, Röhl A, Lünnemann P, et al. A Short Tandem Repeat-Based Phylogeny for the Human Y Chromosome. *American Journal of Human Genetics*. 2000;67(1):182–196.

36. Macaulay VA, Richards MB, Forster P, et al. mtDNA Mutation Rates-No Need to Panic. *Am. J. Hum. Genet.* 1997;61(4):983-986.

37. Heyer E, Zietkiewicz E, Rochowski A, et al. Phylogenetic and Familial Estimates of Mitochondrial Substitution Rates: Study of Control Region Mutations in Deep-Rooting Pedigrees. *Am. J. Hum. Genet.* 2001;69(5):1113-1126.

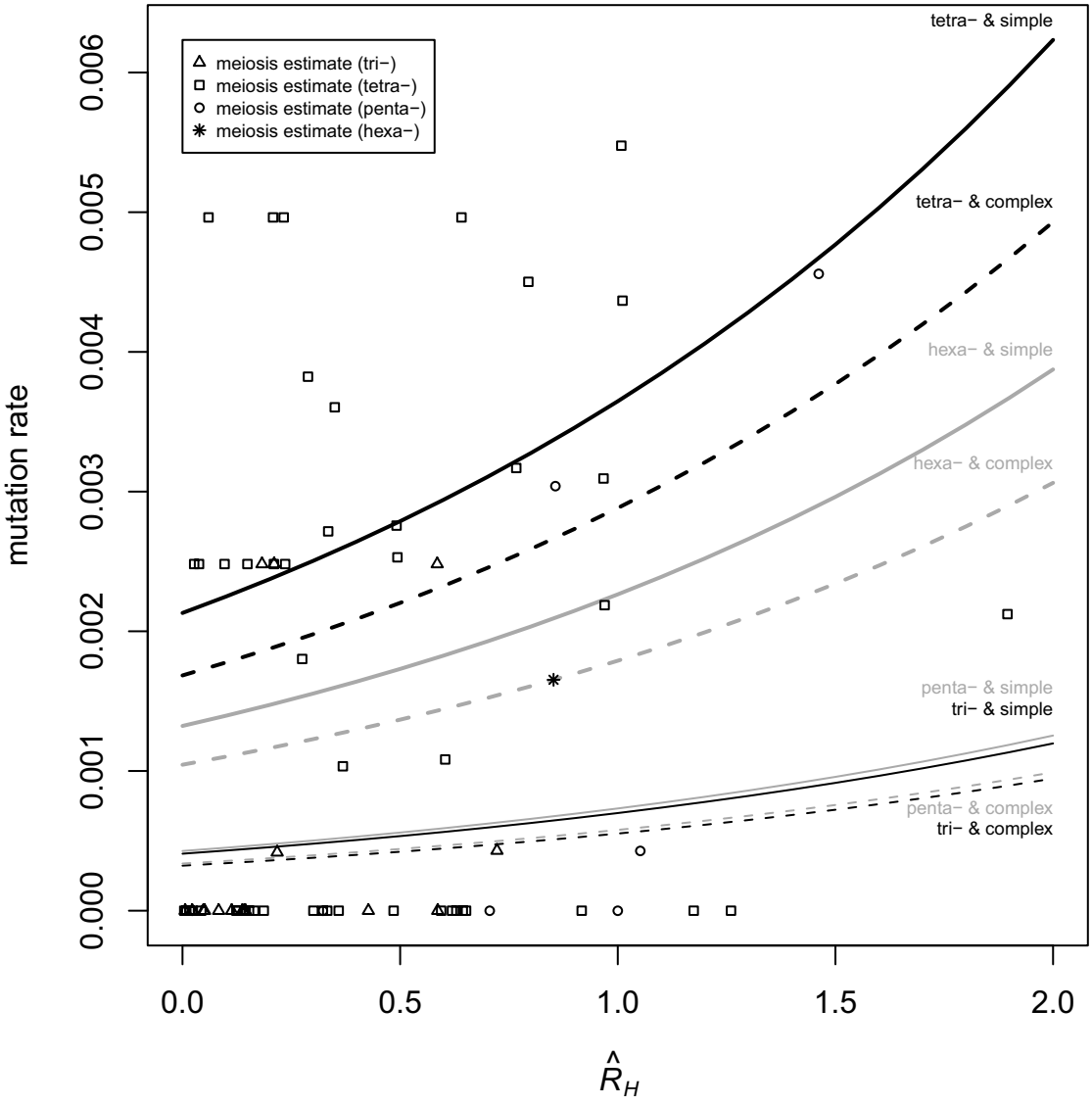
38. Forster P, Kayser M, Meyer E, et al. Phylogenetic resolution of complex mutational features at Y-STR DYS390 in aboriginal Australians and Papuans. *Mol. Biol. Evol.* 1998;15(9):1108–1114.

39. Nebel A, Filon D, Hohoff C, et al. Haplogroup-specific deviation from the stepwise mutation model at the microsatellite loci DYS388 and DYS392. *Eur. J. Hum. Genet.* 2001;9:22–26.

Figure 1. Mutation rate estimates (measured in mutations per generation) from meiosis for 80 Y-STR loci (points) and prediction from logistic regression for the eight categories of loci defined by motif size and repeat structure (lines). Continuous lines represent the predictions for loci with a simple repeat structure and dashed lines for complex loci. Thick black lines are used for the predictions of tetranucleotide loci, thick grey lines for hexa-loci, thin black lines for penta- loci and thin grey lines for tri- loci. The logistic regression model ($L\mu = - 6.863 + 0.539\,\hat{R}_H - 1.176M_{tri} + 0.478M_{tetra} - 1.130M_{penta} + 0.236S_{simple} +$ error, see supplementary table S3 for coefficient p-values) gives the relationship between the logit of mutation rate ($L\mu$) and the predictive variables $\hat{R}_H$ (population relative mutation rate estimated using homozigosity), *M* (motif size: tri-, tetra-, penta- and hexanucleotide classes) and *S* (repeat structure: simple or complex). The model shows that $L\mu$ increases with $\hat{R}_H$ and depends on repeat size (highest for tetranucleotide loci followed by hexa-, penta- and tri- in this order) and on the complexity of the loci (higher for simple than for complex loci).

Figure 2. Root of the relative mean squared error (RrelMSE) for mutation rate estimates calculated from meiosis data (filled circles) or from population data (open circles, triangles and squares) at each of 108 simulated loci. RrelMSE for estimates for loci mutating at true mutation rates of (a) 4×10^{-4} mutations per generation, (b) 8×10^{-4}, (c) 1.6×10^{-3} and (d) 3.2×10^{-3}. RrelMSE for regression based estimates also depends on the category the loci belong to: ‘tri’ (triangles), ‘tetra’ (squares) or ‘penta’ (open circles).

568 **Table 1.** Mutation rate estimates (measured in mutations per generation), obtained from
569 combined meiosis data from 29 published studies (listed in supplementary table S2) and
570 predicted from the logistic model, for 110 Y-STR loci.
571
572


meiosis estimate (tri−)
meiosis estimate (tetra−)
meiosis estimate (penta−)
meiosis estimate (hexa−)
tetra− & simple
tetra− & complex
hexa− & simple
hexa− & complex
penta− & simple
tri− & simple
penta− & complex
tri− & complex
mutation rate
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.0
0.5
1.0
1.5
2.0
$\hat{R}_H$

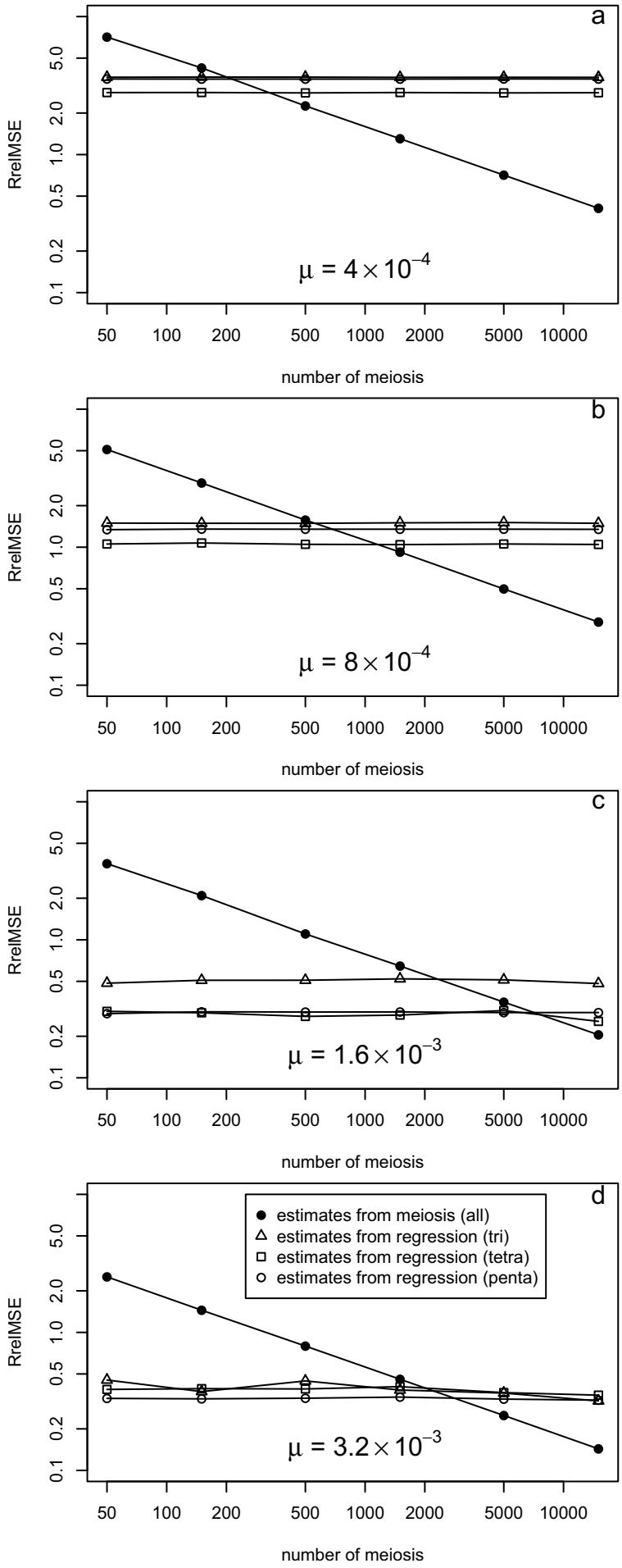

a
b
c
d
RrelMSE
0.1
0.2
0.5
1.0
2.0
5.0
50
100
200
500
1000
2000
5000
10000
number of meiosis
μ = 4 × 10−4
μ = 8 × 10−4
μ = 1.6 × 10−3
μ = 3.2 × 10−3
estimates from meiosis (all)
estimates from regression (tri)
estimates from regression (tetra)
estimates from regression (penta)

Table 1. Mutation rate estimates (measured in mutations per generation), obtained from combined meiosis data from 29 published studies (listed in supplementary table S2) and predicted from the logistic model, for 110 Y-STR loci.

Locus	Mutations in meiosis	Meiosis	$\hat{\mu}_{meiosis}$	95% CI	Motif size[1]	Repeat structure[1]	$\hat{R}_H$ [1]	$\hat{\mu}_{regression}$	95% CI
DYS388[2,3]	*1*	*2394*	*4.177×10^{-4}*	*(2.143×10^{-5}–2.362×10^{-3})*	*tri*	*simple*	*0.218*	*4.587×10^{-4}*	*(2.496×10^{-4} - 8.430×10^{-4})*
DYS426					tri	simple	0.214	4.579×10^{-4}	(2.491×10^{-4} - 8.416×10^{-4})
DYS436[2]					*tri*	*simple*	*0.146*	*4.414×10^{-4}*	*(2.394×10^{-4} - 8.139×10^{-4})*
DYS472	0	403	0	(0–9.442×10^{-3})	tri	simple	0.006	4.094×10^{-4}	(2.204×10^{-4} - 7.601×10^{-4})
DYS476	0	403	0	(0–9.442×10^{-3})	tri	simple	0.051	4.193×10^{-4}	(2.263×10^{-4} - 7.768×10^{-4})
DYS480	0	403	0	(0–9.442×10^{-3})	tri	simple	0.022	4.129×10^{-4}	(2.225×10^{-4} - 7.661×10^{-4})
DYS481	3	403	7.444×10^{-3}	(2.535×10^{-3}–2.166×10^{-2})	tri	simple	5.272	6.937×10^{-3}	(3.163×10^{-3} - 1.514×10^{-2})
DYS485	1	403	2.481×10^{-3}	(1.273×10^{-4}–1.392×10^{-2})	tri	simple	0.585	5.591×10^{-4}	(3.087×10^{-4} - 1.013×10^{-3})
DYS487	1	403	2.481×10^{-3}	(1.273×10^{-4}–1.392×10^{-2})	tri	simple	0.210	4.570×10^{-4}	(2.485×10^{-4} - 8.400×10^{-4})
DYS488	0	403	0	(0–9.442×10^{-3})	tri	simple	0.138	4.394×10^{-4}	(2.382×10^{-4} - 8.105×10^{-4})
DYS490	0	403	0	(0–9.442×10^{-3})	tri	simple	0.143	4.406×10^{-4}	(2.389×10^{-4} - 8.126×10^{-4})
DYS491	0	403	0	(0–9.442×10^{-3})	tri	simple	0.083	4.267×10^{-4}	(2.306×10^{-4} - 7.892×10^{-4})
DYS492	0	403	0	(0–9.442×10^{-3})	tri	simple	0.144	4.410×10^{-4}	(2.391×10^{-4} - 8.132×10^{-4})
DYS494	0	403	0	(0–9.442×10^{-3})	tri	simple	0.048	4.187×10^{-4}	(2.259×10^{-4} - 7.758×10^{-4})
DYS495	0	403	0	(0–9.442×10^{-3})	tri	simple	0.586	5.594×10^{-4}	(3.089×10^{-4} - 1.013×10^{-3})
DYS497	1	403	2.481×10^{-3}	(1.273×10^{-4}–1.392×10^{-2})	tri	simple	0.183	4.502×10^{-4}	(2.445×10^{-4} - 8.286×10^{-4})
DYS617	0	403	0	(0–9.442×10^{-3})	tri	simple	0.427	5.135×10^{-4}	(2.819×10^{-4} - 9.353×10^{-4})
DYS618	0	403	0	(0–9.442×10^{-3})	tri	simple	0.113	4.336×10^{-4}	(2.347×10^{-4} - 8.008×10^{-4})
DYS392	6	13948	4.302×10^{-4}	(1.972×10^{-4}–9.383×10^{-4})	tri	complex	0.722	4.755×10^{-4}	(2.628×10^{-4} - 8.603×10^{-4})
Mean			9.257×10^{-4}					7.974×10^{-4}	
SD			1.931×10^{-3}					1.487×10^{-3}	
DYF406S1					tetra	simple	1.484	4.728×10^{-3}	(3.874×10^{-3} - 5.771×10^{-3})
DYS393 (aka DYS395)	13	12576	1.034×10^{-3}	(6.042×10^{-4}–1.768×10^{-3})	tetra	simple	0.368	2.598×10^{-3}	(2.163×10^{-3} - 3.119×10^{-3})
DYS434[2,3]	*0*	*80*	*0*	*(0–4.582×10^{-2})*	*tetra*	*simple*	*0.359*	*2.584×10^{-3}*	*(2.151×10^{-3} - 3.105×10^{-3})*
DYS435[2,3]	*0*	*161*	*0*	*(0–2.330×10^{-2})*	*tetra*	*simple*	*0.128*	*2.283×10^{-3}*	*(1.874×10^{-3} - 2.780×10^{-3})*
DYS441[2]					*tetra*	*simple*	*1.032*	*3.709×10^{-3}*	*(3.110×10^{-3} - 4.423×10^{-3})*
DYS445[2]					*tetra*	*simple*	*0.272*	*2.467×10^{-3}*	*(2.044×10^{-3} - 2.977×10^{-3})*
DYS453[2]					*tetra*	*simple*	*0.095*	*2.243×10^{-3}*	*(1.837×10^{-3} - 2.738×10^{-3})*
DYS454 (aka DYS639)					*tetra*	*simple*	0.044	2.182×10^{-3}	(1.781×10^{-3} - 2.674×10^{-3})
DYS455[2]					*tetra*	*simple*	*0.008*	*2.140×10^{-3}*	*(1.741×10^{-3} - 2.630×10^{-3})*
DYS456	30	6664	4.502×10^{-3}	(3.155×10^{-3}–6.419×10^{-3})	tetra	simple	0.795	3.266×10^{-3}	(2.748×10^{-3} - 3.881×10^{-3})
DYS458	46	6684	6.882×10^{-3}	(5.164×10^{-3}–9.167×10^{-3})	tetra	simple	1.503	4.777×10^{-3}	(3.908×10^{-3} - 5.838×10^{-3})
DYS460 (aka GATA A7.1)	5	1308	3.823×10^{-3}	(1.634×10^{-3}–8.917×10^{-3})	tetra	simple	0.288	2.488×10^{-3}	(2.064×10^{-3} - 3.000×10^{-3})
DYS461 (aka GATA A7.2)	0	922	0	(4.319×10^{-19}–4.149×10^{-3})	tetra	simple	0.619	2.972×10^{-3}	(2.497×10^{-3} - 3.538×10^{-3})
DYS462[2]					*tetra*	*simple*	*0.490*	*2.774×10^{-3}*	*(2.321×10^{-3} - 3.313×10^{-3})*
DYS505	0	403	0	(0–9.442×10^{-3})	tetra	simple	0.629	2.988×10^{-3}	(2.511×10^{-3} - 3.557×10^{-3})
DYS508	2	403	4.963×10^{-3}	(1.362×10^{-3}–1.791×10^{-2})	tetra	simple	0.641	3.007×10^{-3}	(2.527×10^{-3} - 3.579×10^{-3})
DYS511	1	403	2.481×10^{-3}	(1.273×10^{-4}–1.392×10^{-2})	tetra	simple	0.210	2.386×10^{-3}	(1.969×10^{-3} - 2.890×10^{-3})
DYS522	0	555	0	(0–6.874×10^{-3})	tetra	simple	0.485	2.766×10^{-3}	(2.314×10^{-3} - 3.305×10^{-3})
DYS525	0	403	0	(0–9.442×10^{-3})	tetra	simple	0.187	2.356×10^{-3}	(1.942×10^{-3} - 2.858×10^{-3})
DYS530	0	403	0	(0–9.442×10^{-3})	tetra	simple	0.016	2.150×10^{-3}	(1.750×10^{-3} - 2.640×10^{-3})
DYS531	0	483	0	(0– 7.891×10^{-3})	tetra	simple	0.143	2.301×10^{-3}	(1.891×10^{-3} - 2.800×10^{-3})
DYS533	2	555	3.604×10^{-3}	(9.888×10^{-4}– 1.304×10^{-2})	tetra	simple	0.350	2.572×10^{-3}	(2.140×10^{-3} - 3.091×10^{-3})
DYS537	0	403	0	(0–9.442×10^{-3})	tetra	simple	0.124	2.278×10^{-3}	(1.869×10^{-3} - 2.775×10^{-3})
DYS540	0	403	0	(0–9.442×10^{-3})	tetra	simple	0.153	2.314×10^{-3}	(1.903×10^{-3} - 2.813×10^{-3})

DYS549	1	555	1.802×10^{-3}	$(9.24\times10^{-5}–1.013\times10^{-2})$	tetra	simple	0.275	2.471×10^{-3}	$(2.047\times10^{-3} - 2.981\times10^{-3})$
DYS554	1	403	2.481×10^{-3}	$(1.273\times10^{-4}–1.392\times10^{-2})$	tetra	simple	0.038	2.175×10^{-3}	$(1.774\times10^{-3} - 2.667\times10^{-3})$
DYS556	0	403	0	$(0–9.442\times10^{-3})$	tetra	simple	0.301	2.505×10^{-3}	$(2.079\times10^{-3} - 3.018\times10^{-3})$
DYS565	2	403	4.963×10^{-3}	$(1.362\times10^{-3}–1.791\times10^{-2})$	tetra	simple	0.233	2.415×10^{-3}	$(1.996\times10^{-3} - 2.921\times10^{-3})$
DYS567	0	403	0	$(0–9.442\times10^{-3})$	tetra	simple	0.164	2.328×10^{-3}	$(1.916\times10^{-3} - 2.828\times10^{-3})$
DYS568	0	403	0	$(0–9.442\times10^{-3})$	tetra	simple	0.142	2.300×10^{-3}	$(1.890\times10^{-3} - 2.798\times10^{-3})$
DYS569	0	403	0	$(0–9.442\times10^{-3})$	tetra	simple	0.017	2.150×10^{-3}	$(1.751\times10^{-3} - 2.641\times10^{-3})$
DYS570	7	555	$1.261\times^{10-2}$	$(6.123\times10^{-3}–2.580\times10^{-2})$	tetra	simple	1.264	4.203×10^{-3}	$(3.491\times10^{-3} - 5.059\times10^{-3})$
DYS572	1	403	2.481×10^{-3}	$(1.273\times10^{-4}–1.392\times10^{-2})$	tetra	simple	0.236	2.419×10^{-3}	$(2.000\times10^{-3} - 2.926\times10^{-3})$
DYS573	2	403	4.963×10^{-3}	$(1.362\times10^{-3}–1.791\times10^{-2})$	tetra	simple	0.208	2.383×10^{-3}	$(1.967\times10^{-3} - 2.887\times10^{-3})$
DYS575	1	403	2.481×10^{-3}	$(1.273\times10^{-4}–1.392\times10^{-2})$	tetra	simple	0.027	2.162×10^{-3}	$(1.762\times10^{-3} - 2.653\times10^{-3})$
DYS576	9	555	$1.622\times^{10-2}$	$(8.554\times10^{-3}–3.053\times10^{-2})$	tetra	simple	1.256	4.184×10^{-3}	$(3.477\times10^{-3} - 5.034\times10^{-3})$
DYS578	0	403	0	$(0–9.442\times10^{-3})$	tetra	simple	0.333	2.548×10^{-3}	$(2.118\times10^{-3} - 3.065\times10^{-3})$
DYS579	0	403	0	$(0–9.442\times10^{-3})$	tetra	simple	0.004	2.136×10^{-3}	$(1.737\times10^{-3} - 2.625\times10^{-3})$
DYS580	0	403	0	$(0–9.442\times10^{-3})$	tetra	simple	0.009	2.141×10^{-3}	$(1.742\times10^{-3} - 2.631\times10^{-3})$
DYS583	0	403	0	$(0–9.442\times10^{-3})$	tetra	simple	0.023	2.158×10^{-3}	$(1.758\times10^{-3} - 2.648\times10^{-3})$
DYS636	1	403	2.481×10^{-3}	$(1.273\times10^{-4}–1.392\times10^{-2})$	tetra	simple	0.149	2.309×10^{-3}	$(1.899\times10^{-3} - 2.808\times10^{-3})$
DYS638	1	403	2.481×10^{-3}	$(1.273\times10^{-4}–1.392\times10^{-2})$	tetra	simple	0.097	2.245×10^{-3}	$(1.839\times10^{-3} - 2.740\times10^{-3})$
DYS640	2	403	4.963×10^{-3}	$(1.362\times10^{-3}–1.791\times10^{-2})$	tetra	simple	0.060	2.201×10^{-3}	$(1.798\times10^{-3} - 2.694\times10^{-3})$
DYS641	0	403	0	$(0–9.442\times10^{-3})$	tetra	simple	0.043	2.181×10^{-3}	$(1.780\times10^{-3} - 2.673\times10^{-3})$
DYS19 (aka DYS394)	32	14632	2.187×10^{-3}	$(1.550\times10^{-3}–3.086\times10^{-3})$	tetra	complex	0.970	2.836×10^{-3}	$(2.528\times10^{-3} - 3.182\times10^{-3})$
DYS389I	32	12651	2.529×10^{-3}	$(1.792\times10^{-3}–3.569\times10^{-3})$	tetra	complex	0.494	2.196×10^{-3}	$(1.916\times10^{-3} - 2.517\times10^{-3})$
DYS389B	40	12622	3.169×10^{-3}	$(2.328\times10^{-3}–4.312\times10^{-3})$	tetra	complex	0.767	2.543×10^{-3}	$(2.255\times10^{-3} - 2.867\times10^{-3})$
DYS390	30	14131	2.123×10^{-3}	$(1.488\times10^{-3}–3.029\times10^{-3})$	tetra	complex	1.895	4.661×10^{-3}	$(3.923\times10^{-3} - 5.536\times10^{-3})$
DYS391	38	13995	2.715×10^{-3}	$(1.979\times10^{-3}–3.724\times10^{-3})$	tetra	complex	0.335	2.016×10^{-3}	$(1.735\times10^{-3} - 2.343\times10^{-3})$
DYS437 (aka DYS457)	10	9238	1.082×10^{-3}	$(5.881\times10^{-4}–1.992\times10^{-3})$	tetra	complex	0.604	2.330×10^{-3}	$(2.048\times10^{-3} - 2.649\times10^{-3})$
DYS439 (aka GATA A4)	51	9313	5.476×10^{-3}	$(4.168\times10^{-3}–7.192\times10^{-3})$	tetra	complex	1.008	2.895×10^{-3}	$(2.580\times10^{-3} - 3.248\times10^{-3})$
DYS442[2]					*tetra*	*complex*	*0.250*	*1.926×10^{-3}*	*$(1.644\times10^{-3} - 2.256\times10^{-3})$*
DYS443[2,3]	*0*	*80*	*0*	*$(0–4.582\times10^{-2})$*	*tetra*	*complex*	*0.644*	*2.381×10^{-3}*	*$(2.098\times10^{-3} - 2.701\times10^{-3})$*
DYS444	0	80	0	$(0–4.582\times10^{-2})$	tetra	complex	0.323	2.003×10^{-3}	$(1.722\times10^{-3} - 2.330\times10^{-3})$
DYS449	7	369	1.897×10^{-2}	$(9.219\times10^{-3}–3.863\times10^{-2})$	tetra	complex	3.254	9.642×10^{-3}	$(6.849\times10^{-3} - 1.356\times10^{-2})$
DYS504					tetra	complex	3.183	9.284×10^{-3}	$(6.658\times10^{-3} - 1.293\times10^{-2})$
DYS510[2]					*tetra*	*complex*	*0.664*	*2.407×10^{-3}*	*$(2.124\times10^{-3} - 2.727\times10^{-3})$*
DYS513[2]					*tetra*	*complex*	*0.560*	*2.275×10^{-3}*	*$(1.995\times10^{-3} - 2.596\times10^{-3})$*
DYS520	0	80	0	$(0– 4.582\times10^{-2})$	tetra	complex	0.594	2.318×10^{-3}	$(2.037\times10^{-3} - 2.638\times10^{-3})$
DYS532					tetra	complex	1.687	4.167×10^{-3}	$(3.582\times10^{-3} - 4.847\times10^{-3})$
DYS534					tetra	complex	0.979	2.851×10^{-3}	$(2.541\times10^{-3} - 3.199\times10^{-3})$
DYS544[2]					*tetra*	*complex*	*0.038*	*1.719×10^{-3}*	*$(1.435\times10^{-3} - 2.059\times10^{-3})$*
DYS552[2]					*tetra*	*complex*	*0.971*	*2.838×10^{-3}*	*$(2.529\times10^{-3} - 3.184\times10^{-3})$*
DYS557	0	80	0	$(0– 4.582\times10^{-2})$	tetra	complex	1.260	3.315×10^{-3}	$(2.937\times10^{-3} - 3.740\times10^{-3})$
DYS561[2]					*tetra*	*complex*	*0.151*	*1.827×10^{-3}*	*$(1.544\times10^{-3} - 2.162\times10^{-3})$*
DYS607					tetra	complex	1.481	3.733×10^{-3}	$(3.265\times10^{-3} - 4.268\times10^{-3})$
DYS622[2,3]	*0*	*80*	*0*	*$(0– 4.582\times10^{-2})$*	*tetra*	*complex*	*0.917*	*2.757×10^{-3}*	*$(2.456\times10^{-3} - 3.095\times10^{-3})$*
DYS630[2,3]	*0*	*80*	*0*	*$(0– 4.582\times10^{-2})$*	*tetra*	*complex*	*1.174*	*3.166×10^{-3}*	*$(2.814\times10^{-3} - 3.561\times10^{-3})$*
DYS634					tetra	complex	0.241	1.917×10^{-3}	$(1.635\times10^{-3} - 2.247\times10^{-3})$
DYS635 (aka GATA C4)	23	7434	3.094×10^{-3}	$(2.063\times10^{-3}–4. 638\times10^{-3})$	tetra	complex	0.967	2.832×10^{-3}	$(2.524\times10^{-3} - 3.178\times10^{-3})$
DYS709 (aka DYS516)[2,3]	*0*	*80*	*0*	*$(0– 4.582\times10^{-2})$*	*tetra*	*complex*	*0.651*	*2.390×10^{-3}*	*$(2.108\times10^{-3} - 2.711\times10^{-3})$*
GATA A10[2,3]	*5*	*1145*	*4.367×10^{-3}*	*$(1.867\times10^{-3}–1. 018\times10^{-2})$*	*tetra*	*complex*	*1.011*	*2.899×10^{-3}*	*$(2.584\times10^{-3} - 3.253\times10^{-3})$*
GATA H4	21	7618	2.757×10^{-3}	$(1.804\times10^{-3}–4. 211\times10^{-3})$	tetra	complex	0.492	2.194×10^{-3}	$(1.913\times10^{-3} - 2.515\times10^{-3})$
Mean			2.431×10^{-3}					2.826×10^{-3}	
SD			3.831×10^{-3}					1.309×10^{-3}	
DYS438	4	9339	4.283×10^{-4}	$(1.666\times10^{-4}–1.101\times10^{-3})$	penta	simple	1.052	7.527×10^{-4}	$(3.916\times10^{-4} - 1.446\times10^{-3})$
DYS446	2	658	3.040×10^{-3}	$(8.339\times10^{-4}–1.101\times10^{-2})$	penta	simple	0.857	6.776×10^{-4}	$(3.525\times10^{-4} - 1.302\times10^{-3})$
DYS450					penta	simple	0.176	4.696×10^{-4}	$(2.417\times10^{-4} - 9.123\times10^{-4})$

DYS589	0	403	0	(0–9.442×10^{-3})	penta	simple	0.706	6.248×10^{-4}	(3.246×10^{-4} - 1.202×10^{-3})
DYS590	0	403	0	(0–9.442×10^{-3})	penta	simple	0.023	4.325×10^{-4}	(2.216×10^{-4} - 8.438×10^{-4})
DYS594	0	403	0	(0–9.442×10^{-3})	penta	simple	0.323	5.083×10^{-4}	(2.625×10^{-4} - 9.840×10^{-4})
DYS643[4]	**0**	**555**	**0**	**(0–6.874×10^{-3})**	**penta**	**simple**	**1.000**	**7.320×10^{-4}**	**(3.809×10^{-4} - 1.406×10^{-3})**
YPENTA1					penta	simple	0.517	5.645×10^{-4}	(2.926×10^{-4} - 1.089×10^{-3})
DYS447	3	658	4.559×10^{-3}	(1.552×10^{-3}–1.332×10^{-2})	penta	complex	1.462	7.414×10^{-4}	(3.746×10^{-4} - 1.467×10^{-3})
DYS452[2]					*penta*	*complex*	*0.412*	*4.213×10^{-4}*	*(2.110×10^{-4} - 8.413×10^{-4})*
DYS463					penta	complex	1.307	6.822×10^{-4}	(3.450×10^{-4} - 1.349×10^{-3})
DYS587[2]					*penta*	*complex*	*0.743*	*5.036×10^{-4}*	*(2.538×10^{-4} - 9.991×10^{-4})*
DYS588					penta	complex	0.414	4.217×10^{-4}	(2.112×10^{-4} - 8.421×10^{-4})
DYS593[2]					*penta*	*complex*	*0.487*	*4.387×10^{-4}*	*(2.201×10^{-4} - 8.745×10^{-4})*
DYS645					penta	complex	0.122	3.603×10^{-4}	(1.789×10^{-4} - 7.255×10^{-4})
YPENTA2					penta	complex	0.802	5.198×10^{-4}	(2.622×10^{-4} - 1.030×10^{-3})
Mean			1.147×10^{-3}					5.532×10^{-4}	
SD			1.871×10^{-3}					1.308×10^{-4}	
DYS448	11	6655	1.653×10^{-3}	(9.232×10^{-4}–2.958×10^{-3})	hexa	complex	0.852	1.653×10^{-3}	(9.156×10^{-4} - 2.982×10^{-3})
DYS596					hexa	complex	0.639	1.474×10^{-3}	(8.156×10^{-4} - 2.661×10^{-3})
Mean			na					1.563×10^{-3}	
SD			na					1.268×10^{-4}	

[1] Explanatory variables of the logistic model (supplementary table S3). $\hat{R}_H$, population relative mutation rate based on homozigosity.

[2] Loci with estimates obtained from $\hat{R}'_H$ are marked in italics.

[3] Loci not contributing to the regression because there are no individuals genotyped for them and for the reference locus.

[4] Reference locus is marked in bold.