



HAL
open science

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

Concetta Burgarella, Miguel Navascués

► **To cite this version:**

Concetta Burgarella, Miguel Navascués. Mutation rate estimates for 110 Y-chromosome STRs combining population and father-son pair data.. *European Journal of Human Genetics*, 2011, 19 (1), pp.70-5. 10.1038/ejhg.2010.154 . hal-00584148v2

HAL Id: hal-00584148

<https://hal.science/hal-00584148v2>

Submitted on 1 May 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

3

4 Concetta Burgarella^{1,2} & Miguel Navascués^{1,3*}

5 ¹ CNRS UMR 7625 Écologie et Évolution, École Normale Supérieure, 46 rue d'Ulm

6 75230 Paris (France).

7 ² INIA, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria. Carretera

8 de La Coruña km 7.5, 28040, Madrid (Spain).

9 ³ INRA, UMR CBGP (INRA/IRD/Cirad/Montpellier SupAgro), Campus International de

10 Baillarguet, CS 30016, F-34988 Montferrier-sur-Lez Cedex, France

11

12 * Corresponding author: INRA, UMR CBGP (INRA/IRD/Cirad/Montpellier SupAgro),

13 Campus International de Baillarguet, CS 30016, F-34988 Montferrier-sur-Lez Cedex,

14 France

15 phone: +33(0)4.99.62.33.42

16 fax: +33(0)4.99.62.33.45

17 e-mail: navascues@supagro.inra.fr

18

19

20

21 Running title: Mutation rate estimates for Y-STRs

22 ABSTRACT

23

24 Y-chromosome microsatellites (Y-STRs) are typically used for kinship analysis and
25 forensic identification as well as for inferences on population history and evolution. All
26 applications would greatly benefit from reliable locus-specific mutation rates, to improve
27 forensic probability calculations and interpretations of diversity data. However, estimates of
28 mutation rate from father-son transmissions are available for few loci and have large
29 confidence intervals, due to the small number of meioses usually observed. By contrast,
30 population data exist for many more Y-STRs, holding unused information about their
31 mutation rates. To incorporate single locus diversity information into Y-STR mutation rate
32 estimation, we performed a meta-analysis using pedigree data for 80 loci and individual
33 haplotypes for 110 loci, from 29 and 93 published studies respectively. By means of
34 logistic regression we found that relative genetic diversity, motif size and repeat structure
35 explain the variance of observed rates of mutations from meiosis. This model allowed us to
36 predict locus-specific mutation rates (mean predicted mutation rate 2.12×10^{-3} , SD =
37 1.58×10^{-3}), including estimates for 30 loci lacking meiosis observations and 41 with a
38 previous estimate of zero. These estimates are more accurate than meiosis based estimates
39 when a small number of meioses is available. We argue that our methodological approach,
40 by taking into account locus diversity, could be also adapted to estimate population or
41 lineage specific mutation rates. Such adjusted estimates would represent valuable
42 information for selecting the most reliable markers for a wide range of applications.

43

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

45 glm

46 INTRODUCTION

47

48 Around four hundred microsatellite markers from the Y human chromosome have been
49 made available to date (e.g.¹), with important applications in forensic analyses as well as in
50 genealogy research. However, reliable locus-specific mutation rates are needed to carefully
51 choose loci to minimize the error rate in kinship analysis and sample identification² while
52 obtaining the maximum discriminatory power (e.g.³⁻⁵). Also in population genetics and
53 evolutionary studies, correct inferences on the timing of major demographic events, the age
54 of the most common ancestor, as well as dating Y-lineages and tracing disease evolution
55 are based on the knowledge of mutation rates (e.g.⁶⁻⁸).

56

57 Population genetic theory predicts the genetic diversity of loci in function of their mutation
58 rates (μ) and the effective size of populations (N). Therefore, it is possible to obtain
59 estimates of the joint parameter $\theta=2N\mu$ from genetic diversity indices. In the case of loci
60 evolving under a stepwise mutation model (SMM, generally assumed for microsatellites) it
61 is possible to use the variance (V) in allele repeat count (i.e. allele size measured in number
62 of repeats) and the ‘homozygosity’ ($H = \sum_{i=1}^k p_i^2$, where k is the number of different alleles
63 in the population and p_i is the frequency of the i^{th} allele; note that the term homozygosity is
64 not biologically meaningful for haploid loci but it will be used through the article for the
65 sake of simplicity) for the estimation of θ using the following relationships⁹:

66
$$\hat{\theta}_V = \hat{V} \quad [1]$$

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 rates¹⁰. 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

89 MATERIALS AND METHODS

90

91 *Source of population data*

92

93 Population data for 110 Y- chromosome microsatellite loci have been collected from 93

94 published works, for a total of 22,165 individual haplotypes (note that each individual was

95 genotyped for a subset of loci and never for all of them). Locus names, sample sizes and

96 references are detailed in Supplementary table S1. Locus nomenclature and allele call have

97 been thoroughly checked to assure congruence across works and to remove duplicate data.

98 Any population data with incongruent allele codes were either made uniform (when

99 information provided by authors made it possible unequivocally) or excluded from analysis.

100 Specifically, data from GATA H4 and GATA H4.1 have been pooled under the name

101 GATA H4.1 by applying the appropriate correction to allele calls^{13,14} and DYS389II has

102 been transformed into DYS389B by subtracting allele size of DYS389I¹⁵. Multi-copy loci

103 as well as single individuals with duplicated or variant alleles were excluded from the

104 analysis. Data sets were chosen in order to obtain a maximum representation of loci and of

105 geographical areas; collection of data stopped when no additional data sets could be found

106 that would add data for new loci or would increase the order of magnitude of the sample

107 size for individuals genotyped for a locus.

108

109 *Source of meiosis (father-son pair) data*

110

111 Direct observations of mutation events from meiosis data (father-son pairs) have been
112 collected for 80 loci among the 110 loci with population data, from 29 published studies
113 (table 1 and supplementary table S2). Confidence intervals from binomial probability
114 distribution were estimated according to Wilson method¹⁶. Mutations assigned to
115 DYS389II were carefully checked to discriminate those actually occurring in the DYS389I
116 fragment from those occurred in the DYS389B fragment. Discrimination was always
117 possible except for data from reference ¹¹, which were excluded for this locus.

118

119 *Statistical analysis*

120

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

127

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

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

145

146 Relative mutation rate ($R = \mu_l / \mu_r$) for each locus l was estimated exclusively from
147 individuals genotyped for both locus l and reference locus r . This ensures that the genetic
148 diversity of the sample of both loci has been influenced by the same demographic history
149 (this allows assuming the same effective population sizes). Thus, in the absence of
150 selection, the differences in genetic diversity can be attributed solely to the mutation
151 process. Moreover, because of the complete linkage of Y-STRs, data for both loci will
152 share also the same exact genealogy (even if a selective process was in action). Because
153 both loci have the same genealogy, estimates of the mutation rate ratios will be more
154 accurate than in unlinked loci whose genealogies would vary largely due to the randomness
155 of the coalescent process (e.g. nuclear STRs compared in reference¹⁷). 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 regression¹⁸) 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 structure¹⁹.
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 from²¹⁻²⁴.
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-R² value
187 (calculated by the maximum likelihood method²⁵). Pseudo-R² measures the amount of
188 variation in the observed mutation rates explained by the model. The reduced model with
189 the highest pseudo-R² 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 R²⁶, using packages *binom*²⁷ for calculation of
195 confidence intervals (CI), *ape*²⁸ for calculation of heterozygosity, and *pst*²⁹ for calculation
196 of pseudo-R². 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, three 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 SimCoal2³⁰ 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

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

223

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

228 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

229 number of simulations, $\hat{\mu}_i$ is the estimated mutation rate in simulation i and μ is the true
230 mutation rate) was calculated for the two types of mutation rate estimates at each of the 108
231 loci.

232

233 RESULTS

234

235 Locus DYS643 was selected as reference locus following the criteria described above.

236 Mutation rates relative to reference locus were estimated from repeat count variance and
237 homozygosity for 86 loci, which were used in the logistic regression model (Table 1).

238 Problems of multicollinearity were found between \hat{R}_V and the mean repeat count (A),

239 between \hat{R}_H and A and between \hat{R}_H CG content in motif (P_{CG}). Thus, we considered three

240 alternative models with a different combinations of non-collinear variables each: model m1

241 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 ($L\mu = \beta_0 + \beta_1 \hat{R}_H + \beta_2 M_{tri} + \beta_3 M_{tetra} + \beta_4 M_{penta} + \beta_5 S_{simple} + \text{error}$; 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 not 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 reference³¹). 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 sample²¹. Lower mutability of

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

337

338 The repeat structure explains very little the observed STR mutation rates (table S3), but it is
339 maintained after model reduction using the AIC. However, the coefficient β for simple loci
340 is positive in the final reduced m1 (table S3), while it is negative when the repeat structure
341 is used as the only explanatory variable (table S6). Thus, the effect of the repeat structure
342 on mutation rate is of difficult interpretation. Previous studies have failed to find a
343 relationship of simple versus complex repeats with genetic diversity among loci^{38,32,22}. These
344 results might be due to the lack of effect of repeat structure. However, our qualitative
345 classification of loci as ‘simple’ or ‘complex’ could be missing essential information of
346 complex loci (i.e. differential length of the homogeneous array or combination of variable
347 and constant repeats²¹) affecting the mutation rate. More precise definitions of the degree
348 of complexity, similar to those used in ref.²¹, could yield different results, but require
349 detailed information on loci not readily available.

350

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

355 number of repeats between two loci can be still considered a good estimator of the ratio of
356 mutation rates even in case of multi-step mutations, provided that deviation from the SMM
357 is similar for both loci (cf. equation 2 in reference¹⁷). Although the same argument is not
358 strictly valid for the estimate of θ from homozygosity, small deviations from the SMM
359 change very little the expected homozygosity (cf. Table I from⁹). Only 14 mutations (3.1%
360 of total) involving multiple repeat units are included in meiosis data; therefore, SMM can
361 be considered a reasonable approximation. In addition, the great congruence in prediction
362 between models m1 (using homozygosity) and m2 (using repeat count variance) suggests
363 that the mutation model violation is not an issue for the analysis (results not shown).

364

365 Some important outcomes derive from the approach proposed, emphasizing the positive
366 impact of including population polymorphism data for the improvement of mutation rate
367 estimates and the identification of loci distinctiveness. First, mutation rate estimates were
368 obtained for 30 loci lacking estimates from meiosis observations. Second, locus-specific
369 values of mutation rates can be obtained, while meiosis-based estimates give often equal
370 values for several loci. Third, estimates can be obtained also for loci with very low
371 mutation rates, for which a large sample of meiosis data is required to obtain a non-zero
372 mutation rate estimate. Regarding this point, this work provides mutation rate estimates for
373 41 loci whose mutation rate estimate from meiosis was zero because no mutations had been
374 observed. Lastly, regression based estimates present lower error than estimates from
375 meiosis when only a 'low' number of meiosis is observed.

376

377 The analysis performed in this work represents a valuable tool for selecting most reliable
378 markers to increase Y-STR set currently applied in forensic and kinship analyses. Also, the
379 choice of adequate mutation rates keeps being an issue of great concern when inferences on
380 human diversity and population history are pursued, as put in evidence in a recent work⁶.
381 To account for the different variability of microsatellite loci, these authors use repeat count
382 variance to obtain recalibrated evolutionary mutation rates for groups of loci. Our
383 approximation allows more detailed results by achieving an adjusted mutation rate for each
384 locus separately. The same methodology could be used to estimate population or lineage
385 specific mutation rates, since different lineages and populations are often characterized by
386 specific allele combinations^{32,33,39} and mutation rate seems to be affected by allele size and
387 structure. Finally, the analysis presented here can be easily automated for a data base,
388 allowing the updating of estimates when new population and meiosis data are incorporated
389 from upcoming studies.

390

391

392 ACKNOWLEDGMENTS

393

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

398

399

400 CONFLICT OF INTEREST STATEMENT

401

402 The authors declare no conflict of interest.

403

404

405 Supplementary information is available at European Journal of Human Genetics' website

406 <http://www.nature.com/ejhg/index.html>

407

408

409 REFERENCES

410

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

413

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

416

417 3. Mulero JJ, Chang CW, Calandro LM, et al. Development and Validation of the
418 AmpF ℓ STR $\text{\textcircled{R}}$ Yfiler TM PCR Amplification Kit: A Male Specific, Single Amplification 17
419 Y-STR Multiplex System. *J. For. Sci.* 2006;51(1):64–75.

420

421 4. Lim S, Xue Y, Parkin E, Tyler-Smith C. Variation of 52 new Y-STR loci in the Y
422 Chromosome Consortium worldwide panel of 76 diverse individuals. *Int. J. Legal Med.*
423 2007;121(2):124–127.

424

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

428

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

432

- 433 7. Zerjal T, Wells RS, Yuldasheva N, Ruzibakiev R, Tyler-Smith C. A Genetic Landscape
434 Reshaped by Recent Events: Y-Chromosomal Insights into Central Asia. *Am. J. Hum.*
435 *Genet.* 2002;71(3):466-482.
436
- 437 8. Xue Y, Zerjal T, Bao W, et al. Male Demography in East Asia: A North-South Contrast
438 in Human Population Expansion Times. *Genetics.* 2006;172(4):2431-2439.
439
- 440 9. Kimmel M, Chakraborty R. Measures of variation at DNA repeat loci under a general
441 stepwise mutation model. *Theor. Popul. Biol.* 1996;50(3):345-367.
442
- 443 10. Xu H, Fu Y. Estimating effective population size or mutation rate with microsatellites.
444 *Genetics.* 2004;166(1):555-563.
445
- 446 11. Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P. Estimating Y chromosome
447 specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum. Mol.*
448 *Genet.* 1997;6(5):799-803.
449
- 450 12. Gusmão L, Sánchez-Diz P, Calafell F, et al. Mutation rates at Y chromosome specific
451 microsatellites. *Hum. Mutat.* 2005;26(6):520-528.
452
- 453 13. Gusmão L, Alves C, Beleza S, Amorim A. Forensic evaluation and population data on
454 the new Y-STRs DYS434, DYS437, DYS438, DYS439 and GATA A10. *Int. J. Legal Med.*
455 2002;116(3):139-147.
456
- 457 14. Mulero JJ, Budowle B, Butler JM, Gusmão L. Nomenclature and allele repeat structure
458 update for the Y-STR locus GATA H4. *J. For. Sci.* 2006;51(3):694-694.
459
- 460 15. Butler JM, Reeder DJ. Short Tandem Repeat DNA Internet DataBase. 2009. Available
461 at: <http://www.cstl.nist.gov/strbase/>.
462
- 463 16. Wilson EB. Probable inference, the law of succession, and statistical inference. *J. Am.*
464 *Stat. Assoc.* 1927;22(158):209-212.
465
- 466 17. Chakraborty R, Kimmel M, Stivers D, Davison L, Deka R. Relative mutation rates at
467 di-, tri-, and tetranucleotide microsatellite loci. *Proceedings of the National Academy of*
468 *Sciences of the United States of America.* 1997;94(3):1041-1046.
469
- 470 18. Hosmer DW, Lemeshow S. *Applied Logistic Regression.* 2nd ed. New York:
471 Chichester, Wiley.; 2000.
472
- 473 19. Brinkmann B, Klitschar M, Neuhuber F, Hühne J, Rolf B. Mutation rate in human
474 microsatellites: influence of the structure and length of the tandem repeat. *Am. J. Hum.*
475 *Genet.* 1998;62(6):1408-1415.
476

- 477 20. Ellegren H. Heterogeneous mutation processes in human microsatellite DNA
478 sequences. *Nat. Genet.* 2000;24(4):400–402.
479
- 480 21. Kayser M, Kittler R, Erler A, et al. A Comprehensive Survey of Human Y-
481 Chromosomal Microsatellites. *Am. J. Hum. Genet.* 2004;74(6):1183-1197.
482
- 483 22. Järve M, Zhivotovsky LA, Rootsi S, et al. Decreased Rate of Evolution in Y
484 Chromosome STR Loci of Increased Size of the Repeat Unit. *PLoS ONE.* 2009;4(9):e7276.
485
- 486 23. Gusmão L, Butler JM, Carracedo A, et al. DNA Commission of the International
487 Society of Forensic Genetics (ISFG): An update of the recommendations on the use of Y-
488 STRs in forensic analysis. *Forensic Sci. Int.* 2006;157(2–3):187–197.
489
- 490 24. Leat N, Ehrenreich L, Benjeddou M, Cloete K, Davison S. Properties of novel and
491 widely studied Y-STR loci in three South African populations. *Forensic Sci. Int.*
492 2007;168(2–3):154–161.
493
- 494 25. Long JS. *Regression Models for Categorical and Limited Dependent Variables.*
495 Thousand Oaks, California, USA: Sage Publications, Inc; 1997:297.
496
- 497 26. R Development Core Team. *R: A Language and Environment for Statistical*
498 *Computing.* Vienna, Austria: R Foundation for Statistical Computing; 2009. Available at:
499 <http://www.R-project.org>.
500
- 501 27. Dorai-Raj S. binom: binomial confidence intervals for several parameterizations
502 (version 1.0-5). 2009.
503
- 504 28. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R
505 language. *Bioinformatics.* 2004;20(2):289–290.
506
- 507 29. Jackman S. *pscl: Classes and Methods for R Developed in the Political Science*
508 *Computational Laboratory, Stanford University.* Stanford, California: Department of
509 Political Science, Stanford University; 2008. Available at: <http://pscl.stanford.edu/>.
510
- 511 30. Laval G, Excoffier L. SIMCOAL 2.0: a program to simulate genomic diversity over
512 large recombining regions in a subdivided population with a complex history.
513 *Bioinformatics.* 2004;20(15):2485–2487.
514
- 515 31. Zhivotovsky LA, Underhill PA, Cinnio lu C, et al. The Effective Mutation Rate at Y
516 Chromosome Short Tandem Repeats, with Application to Human Population-Divergence
517 Time. *Am J Hum Genet.* 2004;74(1):50–61.
518
- 519 32. Carvalho-Silva DR, Santos FR, Hutz MH, Salzano FM, Pena SD. Divergent Human Y-
520 Chromosome Microsatellite Evolution Rates. *J. Mol. Evol.* 1999;49(2):204-214.
521

- 522 33. Dupuy BM, Stenersen M, Egeland T, Olaisen B. Y-chromosomal microsatellite
523 mutation rates: differences in mutation rate between and within loci. *Hum. Mutat.*
524 2004;23(2):117–124.
525
- 526 34. Ge J, Budowle B, Aranda XG, et al. Mutation rates at Y chromosome short tandem
527 repeats in Texas populations. *Forensic Sci. Int.: Genet.* 2009;3(3):179–184.
528
- 529 35. Forster P, Röhl A, Lünemann P, et al. A Short Tandem Repeat-Based Phylogeny for
530 the Human Y Chromosome. *American Journal of Human Genetics.* 2000;67(1):182–196.
531
- 532 36. Macaulay VA, Richards MB, Forster P, et al. mtDNA Mutation Rates-No Need to
533 Panic. *Am. J. Hum. Genet.* 1997;61(4):983-986.
534
- 535 37. Heyer E, Zietkiewicz E, Rochowski A, et al. Phylogenetic and Familial Estimates of
536 Mitochondrial Substitution Rates: Study of Control Region Mutations in Deep-Rooting
537 Pedigrees. *Am. J. Hum. Genet.* 2001;69(5):1113-1126.
538
- 539 38. Forster P, Kayser M, Meyer E, et al. Phylogenetic resolution of complex mutational
540 features at Y-STR DYS390 in aboriginal Australians and Papuans. *Mol. Biol. Evol.*
541 1998;15(9):1108–1114.
542
- 543 39. Nebel A, Filon D, Hohoff C, et al. Haplogroup-specific deviation from the stepwise
544 mutation model at the microsatellite loci DYS388 and DYS392. *Eur. J. Hum. Genet.*
545 2001;9:22–26.

546 **Figure 1.** Mutation rate estimates (measured in mutations per generation) from meiosis for
547 80 Y-STR loci (points) and prediction from logistic regression for the eight categories of
548 loci defined by motif size and repeat structure (lines). Continuous lines represent the
549 predictions for loci with a simple repeat structure and dashed lines for complex loci. Thick
550 black lines are used for the predictions of tetranucleotide loci, thick grey lines for hexa-
551 loci, thin black lines for penta- loci and thin grey lines for tri- loci. The logistic regression
552 model ($L\mu = - 6.863 + 0.539 \hat{R}_H - 1.176M_{tri} + 0.478M_{tetra} - 1.130M_{penta} + 0.236S_{simple} +$
553 error, see supplementary table S3 for coefficient p-values) gives the relationship between
554 the logit of mutation rate ($L\mu$) and the predictive variables \hat{R}_H (population relative
555 mutation rate estimated using homozygosity), M (motif size: tri-, tetra-, penta- and
556 hexanucleotide classes) and S (repeat structure: simple or complex). The model shows that
557 $L\mu$ increases with \hat{R}_H and depends on repeat size (highest for tetranucleotide loci followed
558 by hexa-, penta- and tri- in this order) and on the complexity of the loci (higher for simple
559 than for complex loci).

560

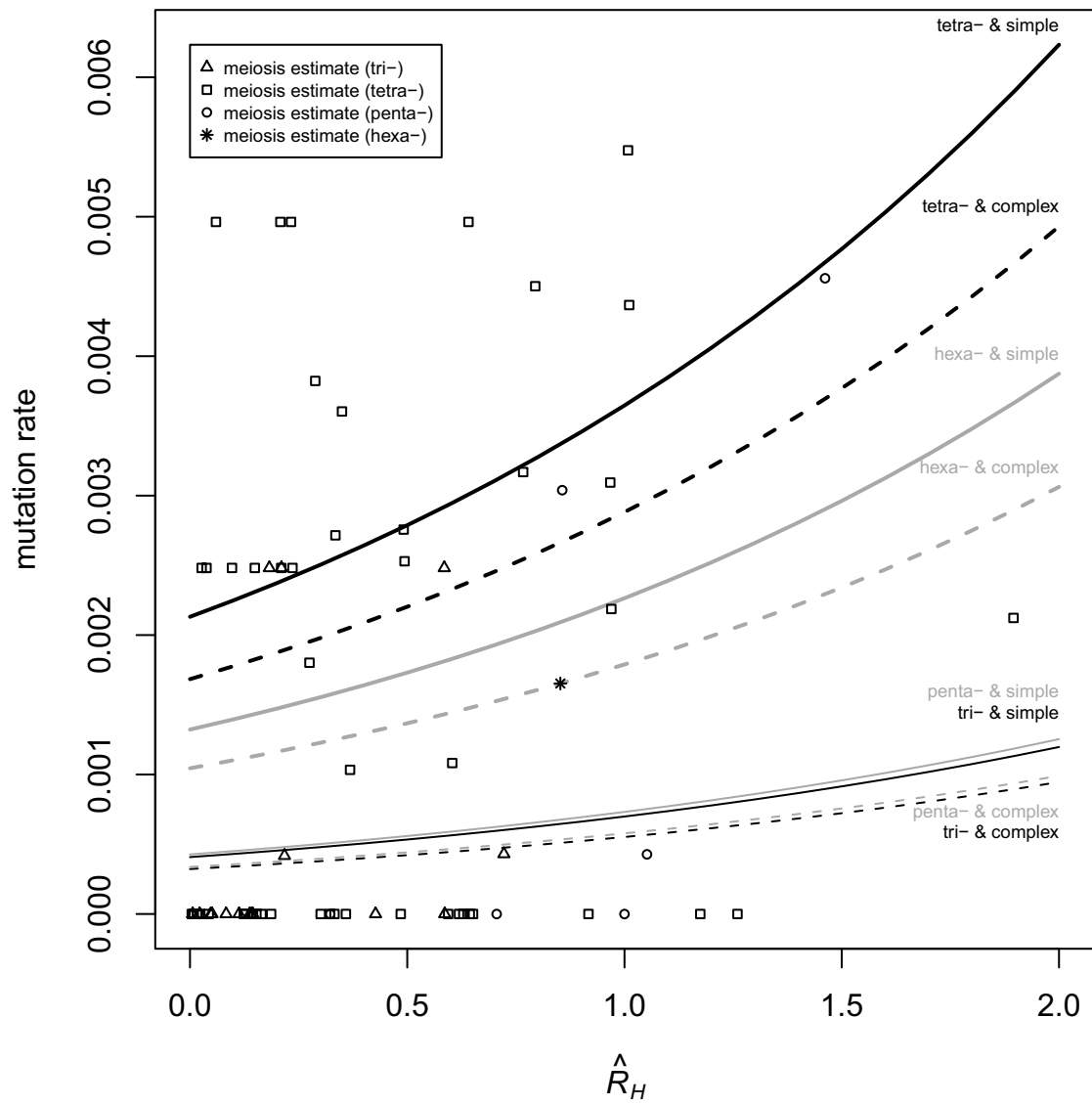
561 **Figure 2.** Root of the relative mean squared error (RrelMSE) for mutation rate estimates
562 calculated from meiosis data (filled circles) or from population data (open circles, triangles
563 and squares) at each of 108 simulated loci. RrelMSE for estimates for loci mutating at true
564 mutation rates of (a) 4×10^{-4} mutations per generation, (b) 8×10^{-4} , (c) 1.6×10^{-3} and (d)
565 3.2×10^{-3} . RrelMSE for regression based estimates also depends on the category the loci
566 belong to: ‘tri’ (triangles), ‘tetra’ (squares) or ‘penta’ (open circles).

567

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



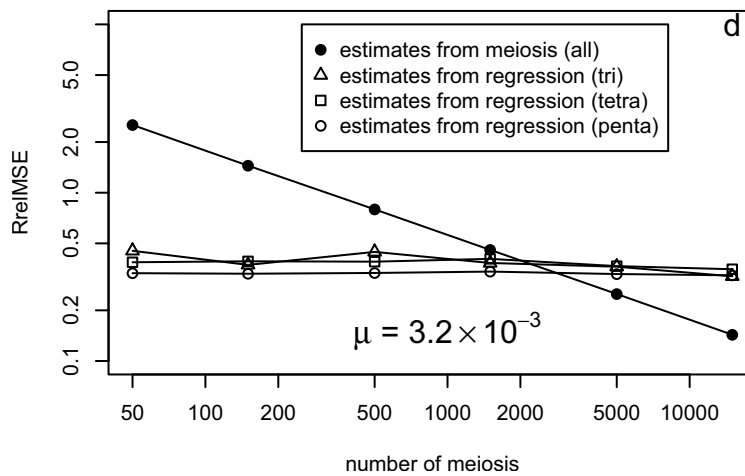
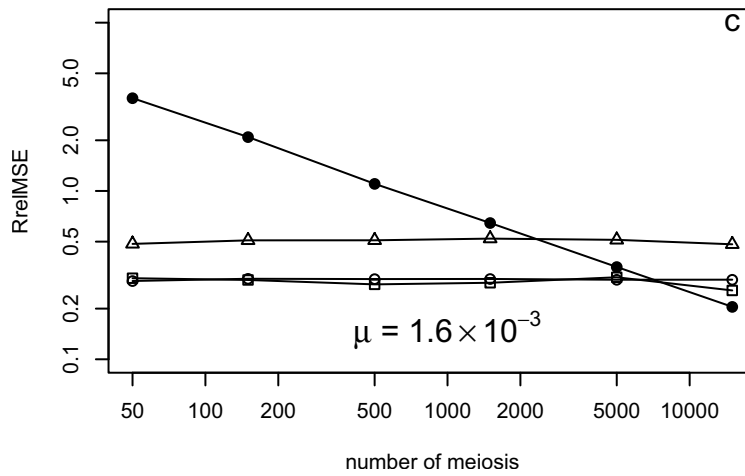
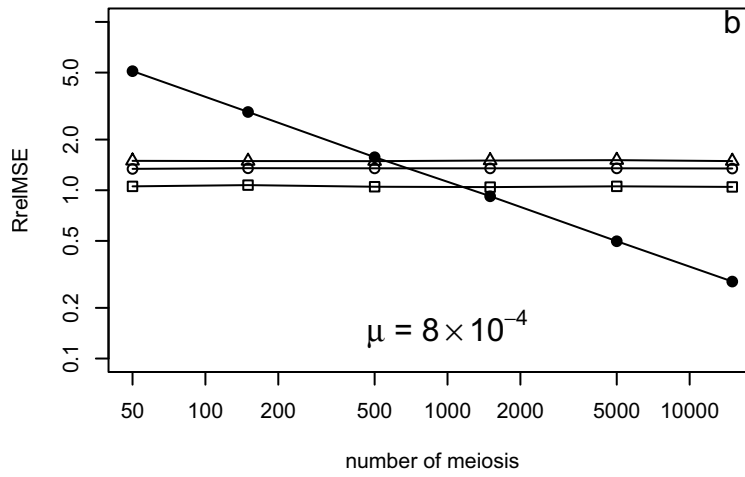
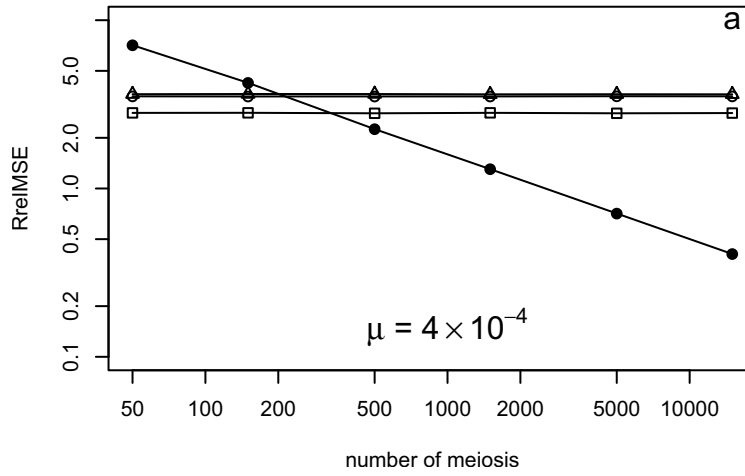


Table 1. Mutation rate estimates (measured in mutations per generation), obtained from combined meiosis data from 29 published studies (listed in supplementary table S2) and predicted from the logistic model, for 110 Y-STR loci.

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

DYS549	1	555	1.802×10^{-3}	$(9.24 \times 10^{-5} - 1.013 \times 10^{-2})$	tetra	simple	0.275	2.471×10^{-3}	$(2.047 \times 10^{-3} - 2.981 \times 10^{-3})$
DYS554	1	403	2.481×10^{-3}	$(1.273 \times 10^{-4} - 1.392 \times 10^{-2})$	tetra	simple	0.038	2.175×10^{-3}	$(1.774 \times 10^{-3} - 2.667 \times 10^{-3})$
DYS556	0	403	0	$(0 - 9.442 \times 10^{-3})$	tetra	simple	0.301	2.505×10^{-3}	$(2.079 \times 10^{-3} - 3.018 \times 10^{-3})$
DYS565	2	403	4.963×10^{-3}	$(1.362 \times 10^{-3} - 1.791 \times 10^{-2})$	tetra	simple	0.233	2.415×10^{-3}	$(1.996 \times 10^{-3} - 2.921 \times 10^{-3})$
DYS567	0	403	0	$(0 - 9.442 \times 10^{-3})$	tetra	simple	0.164	2.328×10^{-3}	$(1.916 \times 10^{-3} - 2.828 \times 10^{-3})$
DYS568	0	403	0	$(0 - 9.442 \times 10^{-3})$	tetra	simple	0.142	2.300×10^{-3}	$(1.890 \times 10^{-3} - 2.798 \times 10^{-3})$
DYS569	0	403	0	$(0 - 9.442 \times 10^{-3})$	tetra	simple	0.017	2.150×10^{-3}	$(1.751 \times 10^{-3} - 2.641 \times 10^{-3})$
DYS570	7	555	1.261×10^{-2}	$(6.123 \times 10^{-3} - 2.580 \times 10^{-2})$	tetra	simple	1.264	4.203×10^{-3}	$(3.491 \times 10^{-3} - 5.059 \times 10^{-3})$
DYS572	1	403	2.481×10^{-3}	$(1.273 \times 10^{-4} - 1.392 \times 10^{-2})$	tetra	simple	0.236	2.419×10^{-3}	$(2.000 \times 10^{-3} - 2.926 \times 10^{-3})$
DYS573	2	403	4.963×10^{-3}	$(1.362 \times 10^{-3} - 1.791 \times 10^{-2})$	tetra	simple	0.208	2.383×10^{-3}	$(1.967 \times 10^{-3} - 2.887 \times 10^{-3})$
DYS575	1	403	2.481×10^{-3}	$(1.273 \times 10^{-4} - 1.392 \times 10^{-2})$	tetra	simple	0.027	2.162×10^{-3}	$(1.762 \times 10^{-3} - 2.653 \times 10^{-3})$
DYS576	9	555	1.622×10^{-2}	$(8.554 \times 10^{-3} - 3.053 \times 10^{-2})$	tetra	simple	1.256	4.184×10^{-3}	$(3.477 \times 10^{-3} - 5.034 \times 10^{-3})$
DYS578	0	403	0	$(0 - 9.442 \times 10^{-3})$	tetra	simple	0.333	2.548×10^{-3}	$(2.118 \times 10^{-3} - 3.065 \times 10^{-3})$
DYS579	0	403	0	$(0 - 9.442 \times 10^{-3})$	tetra	simple	0.004	2.136×10^{-3}	$(1.737 \times 10^{-3} - 2.625 \times 10^{-3})$
DYS580	0	403	0	$(0 - 9.442 \times 10^{-3})$	tetra	simple	0.009	2.141×10^{-3}	$(1.742 \times 10^{-3} - 2.631 \times 10^{-3})$
DYS583	0	403	0	$(0 - 9.442 \times 10^{-3})$	tetra	simple	0.023	2.158×10^{-3}	$(1.758 \times 10^{-3} - 2.648 \times 10^{-3})$
DYS636	1	403	2.481×10^{-3}	$(1.273 \times 10^{-4} - 1.392 \times 10^{-2})$	tetra	simple	0.149	2.309×10^{-3}	$(1.899 \times 10^{-3} - 2.808 \times 10^{-3})$
DYS638	1	403	2.481×10^{-3}	$(1.273 \times 10^{-4} - 1.392 \times 10^{-2})$	tetra	simple	0.097	2.245×10^{-3}	$(1.839 \times 10^{-3} - 2.740 \times 10^{-3})$
DYS640	2	403	4.963×10^{-3}	$(1.362 \times 10^{-3} - 1.791 \times 10^{-2})$	tetra	simple	0.060	2.201×10^{-3}	$(1.798 \times 10^{-3} - 2.694 \times 10^{-3})$
DYS641	0	403	0	$(0 - 9.442 \times 10^{-3})$	tetra	simple	0.043	2.181×10^{-3}	$(1.780 \times 10^{-3} - 2.673 \times 10^{-3})$
DYS19 (aka DYS394)	32	14632	2.187×10^{-3}	$(1.550 \times 10^{-3} - 3.086 \times 10^{-3})$	tetra	complex	0.970	2.836×10^{-3}	$(2.528 \times 10^{-3} - 3.182 \times 10^{-3})$
DYS389I	32	12651	2.529×10^{-3}	$(1.792 \times 10^{-3} - 3.569 \times 10^{-3})$	tetra	complex	0.494	2.196×10^{-3}	$(1.916 \times 10^{-3} - 2.517 \times 10^{-3})$
DYS389B	40	12622	3.169×10^{-3}	$(2.328 \times 10^{-3} - 4.312 \times 10^{-3})$	tetra	complex	0.767	2.543×10^{-3}	$(2.255 \times 10^{-3} - 2.867 \times 10^{-3})$
DYS390	30	14131	2.123×10^{-3}	$(1.488 \times 10^{-3} - 3.029 \times 10^{-3})$	tetra	complex	1.895	4.661×10^{-3}	$(3.923 \times 10^{-3} - 5.536 \times 10^{-3})$
DYS391	38	13995	2.715×10^{-3}	$(1.979 \times 10^{-3} - 3.724 \times 10^{-3})$	tetra	complex	0.335	2.016×10^{-3}	$(1.735 \times 10^{-3} - 2.343 \times 10^{-3})$
DYS437 (aka DYS457)	10	9238	1.082×10^{-3}	$(5.881 \times 10^{-4} - 1.992 \times 10^{-3})$	tetra	complex	0.604	2.330×10^{-3}	$(2.048 \times 10^{-3} - 2.649 \times 10^{-3})$
DYS439 (aka GATA A4)	51	9313	5.476×10^{-3}	$(4.168 \times 10^{-3} - 7.192 \times 10^{-3})$	tetra	complex	1.008	2.895×10^{-3}	$(2.580 \times 10^{-3} - 3.248 \times 10^{-3})$
<i>DYS442²</i>					<i>tetra</i>	<i>complex</i>	<i>0.250</i>	<i>1.926×10^{-3}</i>	<i>$(1.644 \times 10^{-3} - 2.256 \times 10^{-3})$</i>
<i>DYS443^{2,3}</i>	0	80	0	$(0 - 4.582 \times 10^{-2})$	<i>tetra</i>	<i>complex</i>	<i>0.644</i>	<i>2.381×10^{-3}</i>	<i>$(2.098 \times 10^{-3} - 2.701 \times 10^{-3})$</i>
DYS444	0	80	0	$(0 - 4.582 \times 10^{-2})$	tetra	complex	0.323	2.003×10^{-3}	$(1.722 \times 10^{-3} - 2.330 \times 10^{-3})$
DYS449	7	369	1.897×10^{-2}	$(9.219 \times 10^{-3} - 3.863 \times 10^{-2})$	tetra	complex	3.254	9.642×10^{-3}	$(6.849 \times 10^{-3} - 1.356 \times 10^{-2})$
DYS504					tetra	complex	3.183	9.284×10^{-3}	$(6.658 \times 10^{-3} - 1.293 \times 10^{-2})$
<i>DYS510²</i>					<i>tetra</i>	<i>complex</i>	<i>0.664</i>	<i>2.407×10^{-3}</i>	<i>$(2.124 \times 10^{-3} - 2.727 \times 10^{-3})$</i>
<i>DYS513²</i>					<i>tetra</i>	<i>complex</i>	<i>0.560</i>	<i>2.275×10^{-3}</i>	<i>$(1.995 \times 10^{-3} - 2.596 \times 10^{-3})$</i>
DYS520	0	80	0	$(0 - 4.582 \times 10^{-2})$	tetra	complex	0.594	2.318×10^{-3}	$(2.037 \times 10^{-3} - 2.638 \times 10^{-3})$
DYS532					tetra	complex	1.687	4.167×10^{-3}	$(3.582 \times 10^{-3} - 4.847 \times 10^{-3})$
DYS534					tetra	complex	0.979	2.851×10^{-3}	$(2.541 \times 10^{-3} - 3.199 \times 10^{-3})$
<i>DYS544²</i>					<i>tetra</i>	<i>complex</i>	<i>0.038</i>	<i>1.719×10^{-3}</i>	<i>$(1.435 \times 10^{-3} - 2.059 \times 10^{-3})$</i>
<i>DYS552²</i>					<i>tetra</i>	<i>complex</i>	<i>0.971</i>	<i>2.838×10^{-3}</i>	<i>$(2.529 \times 10^{-3} - 3.184 \times 10^{-3})$</i>
DYS557	0	80	0	$(0 - 4.582 \times 10^{-2})$	tetra	complex	1.260	3.315×10^{-3}	$(2.937 \times 10^{-3} - 3.740 \times 10^{-3})$
<i>DYS561²</i>					<i>tetra</i>	<i>complex</i>	<i>0.151</i>	<i>1.827×10^{-3}</i>	<i>$(1.544 \times 10^{-3} - 2.162 \times 10^{-3})$</i>
DYS607					tetra	complex	1.481	3.733×10^{-3}	$(3.265 \times 10^{-3} - 4.268 \times 10^{-3})$
<i>DYS622^{2,3}</i>	0	80	0	$(0 - 4.582 \times 10^{-2})$	<i>tetra</i>	<i>complex</i>	<i>0.917</i>	<i>2.757×10^{-3}</i>	<i>$(2.456 \times 10^{-3} - 3.095 \times 10^{-3})$</i>
<i>DYS630^{2,3}</i>	0	80	0	$(0 - 4.582 \times 10^{-2})$	<i>tetra</i>	<i>complex</i>	<i>1.174</i>	<i>3.166×10^{-3}</i>	<i>$(2.814 \times 10^{-3} - 3.561 \times 10^{-3})$</i>
DYS634					tetra	complex	0.241	1.917×10^{-3}	$(1.635 \times 10^{-3} - 2.247 \times 10^{-3})$
DYS635 (aka GATA C4)	23	7434	3.094×10^{-3}	$(2.063 \times 10^{-3} - 4.638 \times 10^{-3})$	tetra	complex	0.967	2.832×10^{-3}	$(2.524 \times 10^{-3} - 3.178 \times 10^{-3})$
<i>DYS709 (aka DYS516)^{2,3}</i>	0	80	0	$(0 - 4.582 \times 10^{-2})$	<i>tetra</i>	<i>complex</i>	<i>0.651</i>	<i>2.390×10^{-3}</i>	<i>$(2.108 \times 10^{-3} - 2.711 \times 10^{-3})$</i>
<i>GATA A10^{2,3}</i>	5	1145	4.367×10^{-3}	$(1.867 \times 10^{-3} - 1.018 \times 10^{-2})$	<i>tetra</i>	<i>complex</i>	<i>1.011</i>	<i>2.899×10^{-3}</i>	<i>$(2.584 \times 10^{-3} - 3.253 \times 10^{-3})$</i>
GATA H4	21	7618	2.757×10^{-3}	$(1.804 \times 10^{-3} - 4.211 \times 10^{-3})$	tetra	complex	0.492	2.194×10^{-3}	$(1.913 \times 10^{-3} - 2.515 \times 10^{-3})$
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}	$(1.666 \times 10^{-4} - 1.101 \times 10^{-3})$	penta	simple	1.052	7.527×10^{-4}	$(3.916 \times 10^{-4} - 1.446 \times 10^{-3})$
DYS446	2	658	3.040×10^{-3}	$(8.339 \times 10^{-4} - 1.101 \times 10^{-2})$	penta	simple	0.857	6.776×10^{-4}	$(3.525 \times 10^{-4} - 1.302 \times 10^{-3})$
DYS450					penta	simple	0.176	4.696×10^{-4}	$(2.417 \times 10^{-4} - 9.123 \times 10^{-4})$

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)
DYS643⁴	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 ⁻³)	penta	complex	1.462	7.414×10 ⁻⁴	(3.746×10 ⁻⁴ - 1.467×10 ⁻³)
<i>DYS452²</i>					<i>penta</i>	<i>complex</i>	<i>0.412</i>	<i>4.213×10⁻⁴</i>	<i>(2.110×10⁻⁴ - 8.413×10⁻⁴)</i>
DYS463					penta	complex	1.307	6.822×10 ⁻⁴	(3.450×10 ⁻⁴ - 1.349×10 ⁻³)
<i>DYS587²</i>					<i>penta</i>	<i>complex</i>	<i>0.743</i>	<i>5.036×10⁻⁴</i>	<i>(2.538×10⁻⁴ - 9.991×10⁻⁴)</i>
DYS588					penta	complex	0.414	4.217×10 ⁻⁴	(2.112×10 ⁻⁴ - 8.421×10 ⁻⁴)
<i>DYS593²</i>					<i>penta</i>	<i>complex</i>	<i>0.487</i>	<i>4.387×10⁻⁴</i>	<i>(2.201×10⁻⁴ - 8.745×10⁻⁴)</i>
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 homozygosity.

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.