Genet. Res., Camb. (2000), **75**, *pp.* 167–177. *With 5 figures.* Printed in the United Kingdom © 2000 Cambridge University Press

Bayesian estimation of range for microsatellite loci

FEDERICO M. STEFANINI[1] AND MARCUS W. FELDMAN[2]*
[1] *Department of Statistics 'G. Parenti', Università degli Studi di Firenze, Florence, Italy*
[2] *Department of Biological Sciences, Stanford University, Stanford, CA 94035, USA*

(*Received 8 February 1999 and in revised form 23 June 1999*)

Summary

Microsatellite loci have become important in population genetics because of their high level of polymorphism in natural populations, very frequent occurrence throughout the genome, and apparently high mutation rate. Observed repeat numbers (alleles size) in natural populations and expectations based on computer simulations suggest that the range of repeat numbers at a microsatellite locus is restricted. This range is a key parameter that should be properly estimated in order to proceed with calculations of divergence times in phylogenetic studies and to better investigate the within- and between-population variability. The 'plug-in' estimate of range based on the minimum and maximum value observed in a sample is not satisfactory because of the relatively large number of alleles in comparison with typical sample sizes. In this paper, a set of data from 30 dinucleotide microsatellite loci is analysed under the assumption of independence among loci. Bayesian inference on range for one locus is obtained by assuming that constraints on range values exist as sharp bounds. Closed-form calculations and robustness revealed by our analysis suggest that the proposed Bayesian approach might be routinely used by researchers to classify microsatellite loci according to the estimated value of their allelic range.

1. Introduction

Microsatellite loci, also called VNTR (Variable Number of Tandem Repeats) loci, are characterized by the repetition in tandem of a fundamental motif comprising a short sequence of nucleotides (up to 5 bp). Alleles at a microsatellite locus differ in the number of repetitions of the fundamental motif; this number may be called the repeat number or allele size. A non-negative integer can therefore be associated with each allele.

Microsatellite polymorphism seems to be due mainly to the interaction of mutation and genetic drift, although selection may also be important in some trinucleotide loci (Valdes *et al.*, 1993). The stepwise mutation model (Ohta & Kimura, 1973; Moran, 1975) has been used to characterize the dynamics of population mean, variance, kurtosis and skewness (Goldstein *et al.*, 1995; Slatkin, 1995; Zhivotovsky & Feldman, 1995) under the assumption that the range of repeat numbers is infinite.

Several authors have pointed out that constraints on the range of allele sizes (repeat numbers) should be taken into account in phylogenetic studies (e.g. Valdes *et al.*, 1993; Bowcock *et al.*, 1994; Goldstein *et al.*, 1995; Garza *et al.*, 1995; Nauta & Weissing, 1996). Proper estimates of range are important not only because some evolutionary models require knowledge of this parameter, but also as an aid in choosing those loci that are most suitable for phylogenetic investigation (Pollock *et al.*, 1998).

Two main approaches have been taken to incorporate range constraints into evolutionary studies of microsatellite loci. The first uses a model proposed by Garza *et al.* (1995) and further analysed by Zhivotovsky *et al.* (1997). It retains the assumption of an infinite range but takes the stepwise mutation process to be biased towards a fixed central value, so that alleles are more likely to mutate towards a target size in proportion to their difference from it. The effect of this bias is to constrain the range of repeat numbers.

Second, Nauta & Weissing (1996) and Feldman *et al.* (1997) studied a model for microsatellites in which

* Corresponding author.

two allele sizes behave as reflecting barriers, so that an unbiased stepwise process acts on all but these two alleles, the upper and lower allele size boundaries, for which the direction of mutation is unique. Further investigation of this reflecting-boundary mutation model was made by Pollock *et al.* (1998), who also proposed methods of estimating range in the presence of variation across loci in ranges and mutation rates. The estimate of range is based on the family of uniform distributions with two parameters representing the minimum and maximum repeat number. The choice of that family is motivated by the necessity of easy closed-form computation. Moreover, the asymptotic distribution of repeat number under the model assumed for the population dynamics is uniform. The family of uniform distributions on a parameterized range is, however, unlikely to accurately describe a general locus in any natural population.

In this note, we use a more formal approach to the estimation of constraints on allele size, in which Greek letters are used to indicate population parameters. We assume that the repeat number at a microsatellite locus is defined on the bounded set of integers $\Omega_{\alpha,\beta}$ (modified, Feldman *et al.*, 1997; Pollock *et al.*, 1998):

$$\Omega_{\alpha,\beta} = \{\alpha, \alpha+1, \ldots, \beta-1, \beta : 0 \leqslant \alpha \leqslant \beta < \infty\}, \qquad (1)$$

where α and β, respectively the minimum and maximum repeat number, are in general unknown.

The set $\Omega_{\alpha,\beta}$ contains the repeat numbers that may be sampled at one locus. Plug-in estimates of α, β and range $\rho = \beta - \alpha$ are obtained from a sample of size n and realized values $\underline{x} = (x_1, \ldots, x_n)$ as

$$\hat{\rho} = \hat{\beta} - \hat{\alpha} = x^{(n)} - x^{(1)}, \qquad (2)$$

where $x^{(n)}$ and $x^{(1)}$ are, respectively, realizations of the last and first order statistics $X^{(n)}$ and $X^{(1)}$ (Casella & Berger, 1990, p. 229).

It is well known that the estimator $T_0(\underline{X}) = X^{(n)} - X^{(1)}$ is biased downward, because the estimated value $\hat{\rho}$ is equal to ρ only if both α and β are among the sampled values, and the probability of this event is less than one. For small but reasonable probability values on the maximum and minimum repeat number, and for typical sample sizes of 20–100, Stefanini (1997) numerically investigated the features of T_0 in a subclass of microsatellite loci and found that T_0 is biased strongly downward. The severity of the bias increases as the probability values of α and β decrease, and for long-tailed distributions.

Stefanini (1997) also proposed an improved estimator T_1 of ρ for a restricted class of microsatellite loci by applying an approximate correction for bias that depends on the first four estimated moments of the distribution. This approach is not general because it rests mainly on the numerical characterization of correction terms, locus by locus.

Here we begin with a formalization of the estimation problem and with a description of the features for a generic microsatellite locus. Bayesian calculations are developed to obtain the posterior distribution of range for repeat numbers at one locus, and a point estimate is proposed as the mode of the posterior distribution. A Bayesian test of hypothesis is performed to decide whether the sample range should be considered as the population range. The proposed model is applied to published data on 30 dinucleotide microsatellites (Bowcock *et al.*, 1994). The numerical evaluation of robustness against different choices of model components suggests that our method might be used routinely by researchers to classify microsatellite loci according to the estimated values of their ranges. Credibility sets based on the posterior distributions for range values may also be obtained to investigate the sensitivity of some phylogenetic models to the choice of the range parameter.

2. Materials and methods

(i) *Families of distributions for microsatellite loci*

The minimum and maximum repeat numbers α, β are assumed to be specific features of the microsatellite locus. The data we use were sampled from a number of human populations. We shall assume that, for a given locus, α and β are the same in all populations sampled. This allows us to pool data for one locus taken from all the populations.

The discrete random variable X represents a repeat number sampled from a reference (human) population at a given time. The most general shape for the discrete distribution of X is obtained using a parameter π_i for each repeat number i in a large set of values that includes α and β. In other words, the two boundaries are finite and relatively small, $0 \leqslant \alpha \leqslant \beta \leqslant k$, where k is a constant conveniently chosen as 1000 for the class of dinucleotide microsatellites in our analysis. We denote by $\underline{\pi}$ the vector $(\pi_0, \pi_1, \ldots, \pi_{1000})$. The probability mass function of X refers to the probability of sampling a repeat number, $P_r(X = i) = \pi_i$, from the reference population at a given locus, that is

$$p(X; \underline{\pi}, \alpha, \beta) = \sum_{i=\alpha}^{\beta} \pi_i . \mathbf{I}_{\{i\}}(x), \qquad (3)$$

where $\sum_{i=\alpha}^{\beta} \pi_i = 1$, $\pi_\alpha > 0$, $\pi_\beta > 0$, $\pi_i \geqslant 0$ for $\alpha < i < \beta$ with $0 \leqslant \alpha \leqslant \beta \leqslant 1000$ and $\mathbf{I}_{\{i\}}(x) = 1$ if $x = i$ and zero otherwise.

The goal of the analysis is to estimate the range $\rho = \beta - \alpha$, that is a function of the bounds β and α. Estimates of ρ, β and α based on a sample of size n are indicated respectively as $\hat{\rho}$, $\hat{\beta}$ and $\hat{\alpha}$.

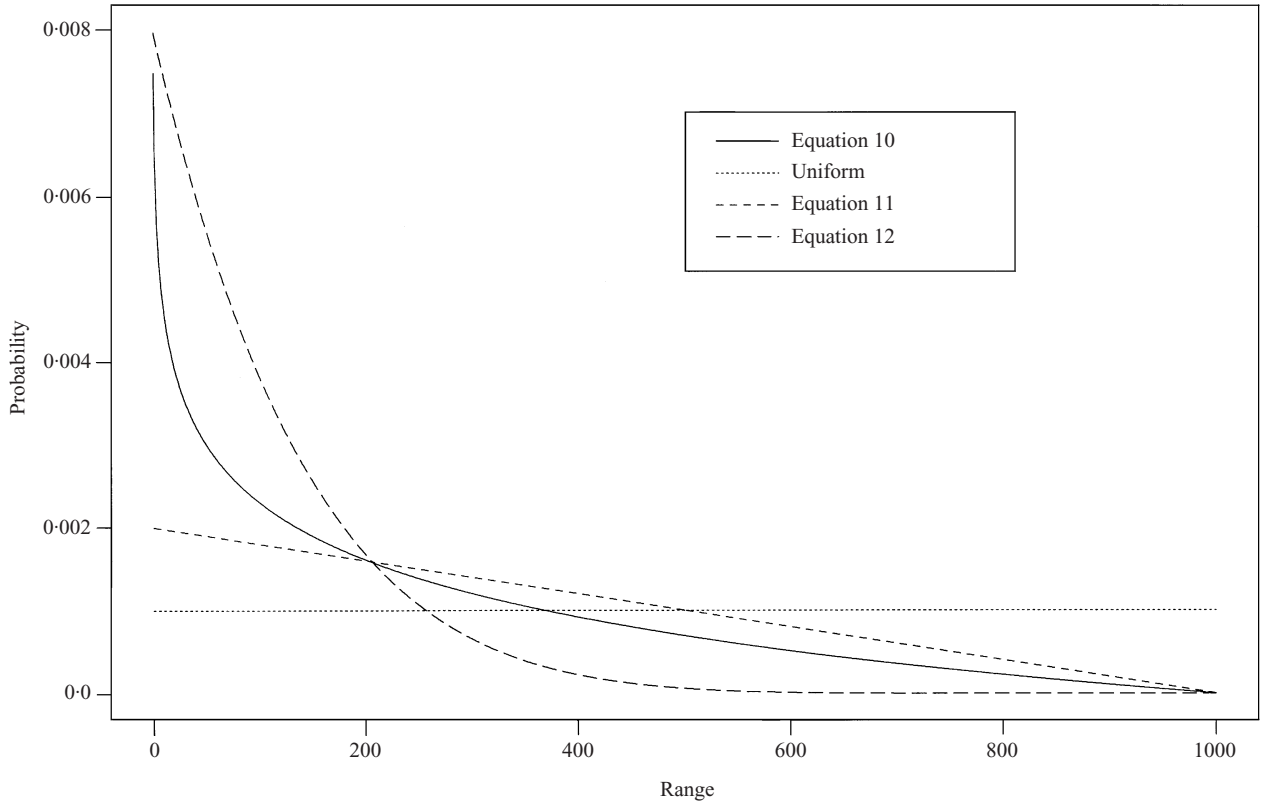

Fig. 1. Plots of the four discrete prior distributions described in the text. All but the uniform prior distributions are labelled with the equation number.

(ii) *Inference about unknown parameters*

For inferences about $\underline{\pi}$, α and β, and $\rho = \beta - \alpha$, we use the Bayesian paradigm (Berger, 1985; O'Hagan, 1994; Bernardo & Smith, 1994).

The proposed Bayesian model has two main components: the likelihood function $p(\underline{x} \mid \underline{\pi}, \alpha, \beta)$ and the prior distribution $p(\underline{\pi}, \alpha, \beta)$. The parameters $\underline{\pi}, \alpha, \beta$ are considered to be random variables, and initial beliefs about reasonable values of parameters are expressed through the prior distribution $p(\underline{\pi}, \alpha, \beta)$.

Bayesian inference is based on the posterior (or final) distribution of the parameter obtained by means of Bayes' theorem:

$$p(\underline{\pi}, \alpha, \beta \mid \underline{x}) = \frac{p(\underline{x} \mid \underline{\pi}, \alpha, \beta) \,.\, p(\underline{\pi} \mid \alpha, \beta) \,.\, p(\alpha, \beta)}{\sum_{\alpha,\beta} \int p(\underline{x} \mid \underline{\pi}, \alpha, \beta) \,.\, p(\underline{\pi} \mid \alpha, \beta) \,.\, p(\alpha, \beta) \,.\, d\underline{\pi}}, \quad (4)$$

where $p(\underline{\pi} \mid \alpha, \beta) \,.\, p(\alpha, \beta) = p(\underline{\pi}, \alpha, \beta)$ is a hierarchical definition of the prior distribution.

For the present analysis, $\underline{\pi}$ is a vector of nuisance parameters that are required to specify the model, but are not of direct interest. They are integrated out of the likelihood to obtain $p(\underline{x} \mid \alpha, \beta)$. Inferences about the two parameters α, β might be obtained using the marginal posterior distribution $p(\alpha, \beta \mid \underline{x})$ that properly summarizes the overall information available about (α, β). Instead, we focus on the range, and its posterior distribution is calculated by means of the transformation $\rho = \beta - \alpha$.

A point estimate of ρ is an effective summary of the posterior distribution of the range if the variance is small with respect to the goal of the inference. If the variance of the posterior distribution $p(\rho \mid \underline{x})$ is not negligible, a better summary is obtained through the highest posterior density (HPD) region (O'Hagan, 1994, par 2·50 and 2·51). This is a credibility region R obtained by collecting the range values with highest posterior probability so that $\mathbf{P}[\rho \in R \mid \underline{x}]$ is large, say 0·95.

It is also interesting to test the hypotheses that the plug-in estimate of range, i.e. the sample range, is the true unknown value of population range (see O'Hagan, 1994, par 3·53 and 3·54; Berger, 1985, p. 145). According to the features of the stepwise mutation process, further interesting hypotheses involve values of range close to the observed value. That is, sampling a repeat number i increases the belief that $i \pm c$ also belong to the population (c a small integer). The hypotheses $\mathbf{H}^{(i)}$, $i = 0, 1, 2, \ldots, 7$ are defined as:

$$\begin{aligned} &\mathbf{H}^{(0)}: \rho = x^{(n)} - x^{(1)} \\ &\mathbf{H}^{(i)}: \rho = x^{(n)} - x^{(1)} + i, \quad \text{for } i = 1, \ldots, 6 \\ &\mathbf{H}^{(7)}: \rho \geqslant x^{(n)} - x^{(1)} + 7. \end{aligned} \quad (5)$$

Table 1. *Observed locus parameters*

Locus	r	$x^{(n)}$	$x^{(1)}$	n
D13S270	20	99	79	266
D13S126	18	114	96	244
D13S119	28	140	112	284
D13S118	14	201	187	282
D13S125	30	161	131	270
D13S144	18	199	181	288
UTSW1523	10	185	175	260
ACTC	28	99	71	264
D15S171	16	125	109	288
D15S169	22	162	140	286
D13S133	70	189	119	266
D13S137	24	123	99	268
D13S227	32	164	132	260
FES	24	167	143	288
GABRB3	20	201	181	268
D13S192	28	119	91	254
D13S193	22	150	128	274
HLIP	14	175	161	280
D15S98	34	175	141	270
D15S97	30	184	154	284
D15S100	30	133	103	290
D15S101	24	138	114	250
D13S115	18	180	162	258
D15S95	18	150	132	290
D15S108	24	163	139	290
D13S71	18	85	67	282
D15S102	22	118	96	290
D15S117	24	150	126	242
D15S148	20	151	131	282
D15S11	26	264	238	286

The columns from left to right are: locus name, observed range ($r = x^{(n)} - x^{(1)}$), maximum ($x^{(n)}$) and minimum ($x^{(1)}$) of repeat numbers, sample size (n).

Multiple hypothesis testing (Berger, 1985, p. 157) is performed by assessing the posterior probability values of the hypotheses in (5). The range cannot be smaller than the observed sample range; that is the eight hypotheses in (5) represent a partition of the parameter space for the range. Thus, the decision is whether the population range is exactly equal to the sample range, greater by just a few units or substantially greater.

(iii) *Model specification*

The likelihood function $p(\underline{x}\,|\,\underline{\pi}, \alpha, \beta)$ is derived from (3) under the assumption of conditional independence among the random variables X_j, $j = 1, \ldots, n$, as $\prod_{j=1}^{n} p(x_j\,|\,\underline{\pi}, \alpha, \beta)$. The use of sufficient statistics for $\underline{\pi}$ (O'Hagan, 1994, par 3·7) allows the likelihood function to be expressed in terms of counts of repeat numbers, indicated here as $\underline{x} = \underline{n} = (n_\alpha, \ldots, n_\beta)$:

$$p(\underline{x}\,|\,\underline{\pi}, \alpha, \beta) = \frac{n!}{\prod_{i=\alpha}^{\beta} n_i!} \cdot \prod_{i=\alpha}^{\beta} \pi_i^{n_i}. \tag{6}$$

We chose a Dirichlet distribution $p(\underline{\pi}\,|\,\alpha, \beta)$ as a prior:

$$p(\underline{\pi}\,|\,\alpha, \beta) = \frac{\Gamma(\lambda)}{\prod_{i=\alpha}^{\beta} \Gamma(\lambda_i)} \prod_{i=\alpha}^{\beta} \pi_i^{\lambda_i - 1}, \tag{7}$$

with parameters $\lambda = \sum_{i=\alpha}^{\beta} \lambda_i = 1$, and $\lambda_i = 1/(\beta - \alpha + 1)$. The lack of knowledge about probabilities of repeat numbers in the population is reflected by the small value of λ. A few sampled observations may cause a detectable change in shape of the distribution of $\underline{\pi}$ (O'Hagan, 1994, par 10·1–10·6). Moreover, the expectation $\mathbf{E}[\pi_i\,|\,\alpha, \beta]$ of the population frequency is a constant that does not depend on the repeat size, as obtained from theoretical calculations of the equilibrium distributions (Feldman *et al*., 1997).

It may be shown that the integral of the likelihood function with respect to $\underline{\pi}$ is proportional to moments of the nuisance parameters within $\underline{\pi}$:

$$\begin{aligned} p(\underline{x}\,|\,\alpha, \beta) &= \int p(\underline{x}\,|\,\underline{\pi}, \alpha, \beta) \,.\, p(\underline{\pi}\,|\,\alpha, \beta) \,.\, d\underline{\pi} \\ &= \frac{n!}{\prod_{i=\alpha}^{\beta} n_i!} \,.\, \mathbf{E}\left[\prod_{i=\alpha}^{\beta} \pi_i^{n_i}\,|\,\alpha, \beta \right]. \end{aligned} \tag{8}$$

Specification of the model is completed by the choice of the prior distribution $p(\alpha, \beta)$. The moderate state of knowledge about α and β suggests a distribution with a large variance, such as

$$\left.\begin{aligned} p(\alpha) &= \text{uniform}(0, 1000) \\ p(\beta\,|\,\alpha) &= \text{uniform}(\alpha, 1000). \end{aligned}\right\} \tag{9}$$

Equation (9) implicitly defines the prior distribution of the range ρ. From (9) and the equality $p(\alpha, \beta) = p(\beta\,|\,\alpha) \,.\, p(\alpha)$, the prior distribution $p(\rho)$ of $\rho = \beta - \alpha$ is obtained by straightforward calculations:

$$p(\rho) = \frac{1}{1001} \cdot \sum_{i=0}^{1000-\rho} \frac{1}{(1001 - i)} \cdot \mathbf{I}_{\{0, \ldots, 1000\}}(\rho), \tag{10}$$

where $\mathbf{I}_{\{0, \ldots, 1000\}}(\rho) = 1$ if ρ belongs to $\{0, \ldots, 1000\}$, and zero otherwise.

The robustness to the choice in (10) is investigated by assessing how point estimates of ρ vary with the choice among three other prior distributions. The first is a uniform distribution for ρ, which has a larger variance than (10). The second is linear on $\{0, \ldots, 1000\}$, and is characterized by

$$\frac{p(\rho = 0)}{p(\rho = 1000)} = 1001;$$

that is,

$$p(\rho) = \frac{1001 - \rho}{501501} \cdot \mathbf{I}_{\{0, \ldots, 1000\}}(\rho). \tag{11}$$

The third is less dispersed than (10), and is obtained

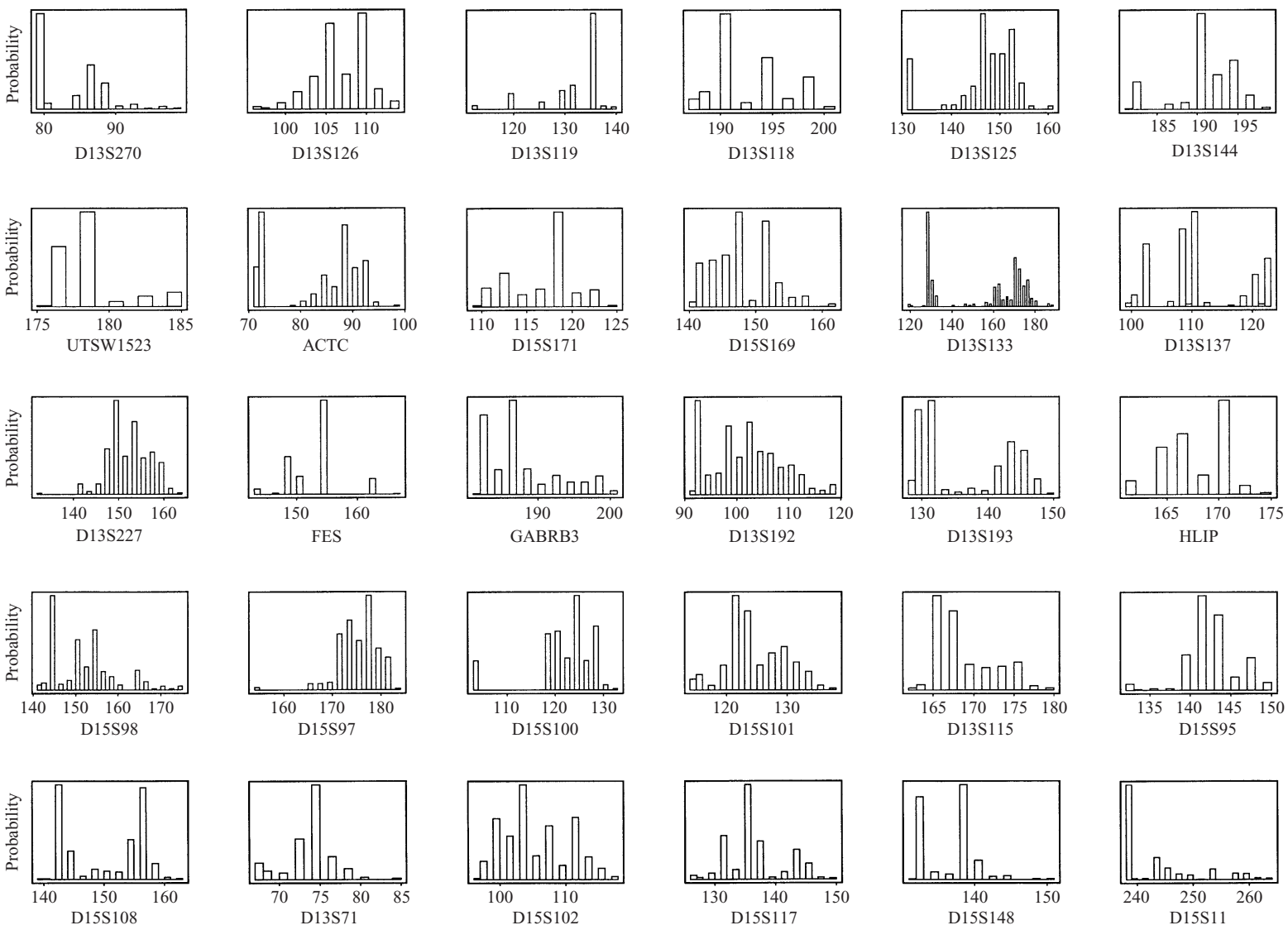

Fig. 2. Histograms of repeat numbers. The name of the locus is reported on the *x*-axis.

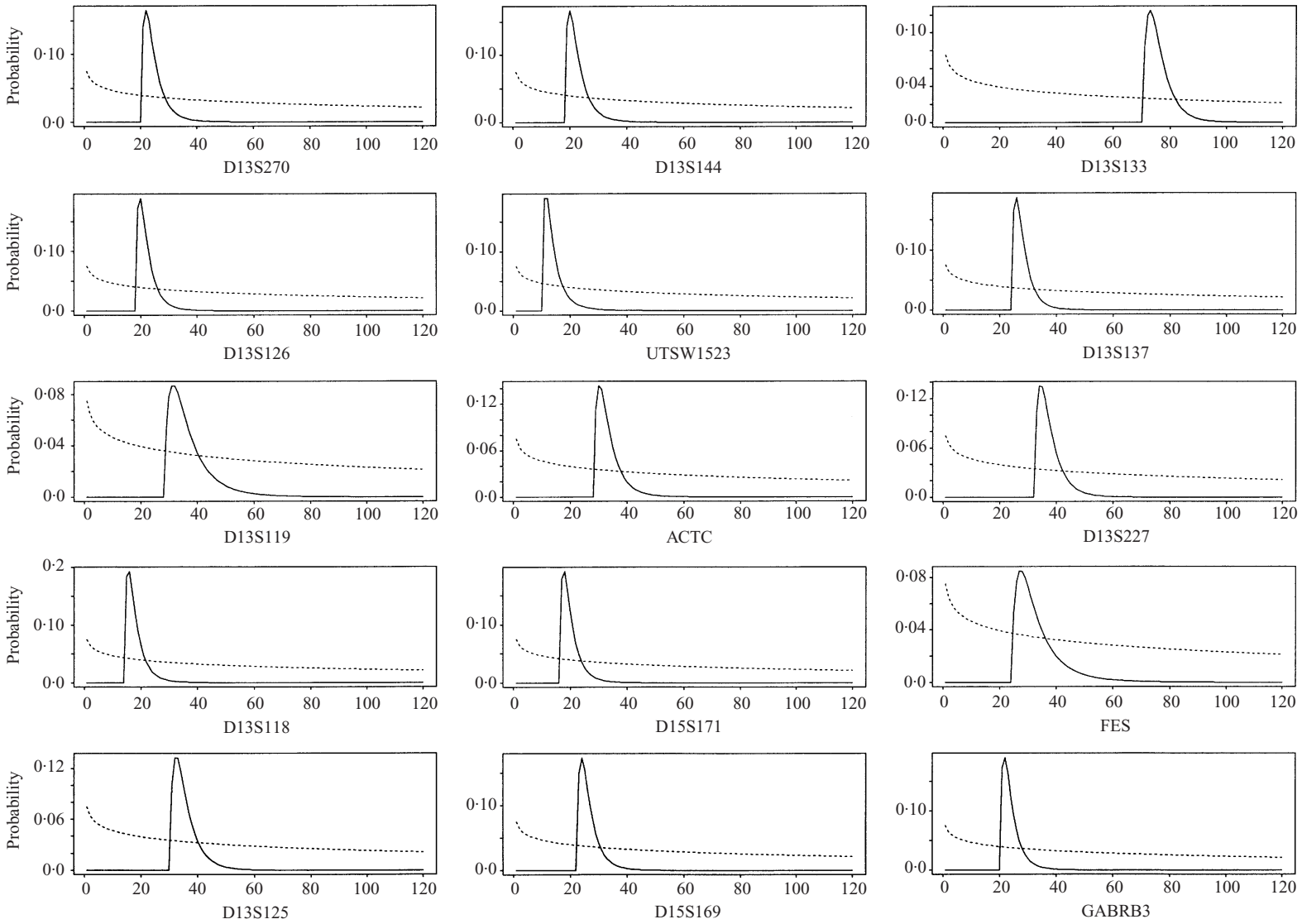

Fig. 3. Posterior distributions of range given the working prior (first 15 loci). The continuous line shows the overall shape of the distribution. The prior distribution for the range is drawn as a dotted line.

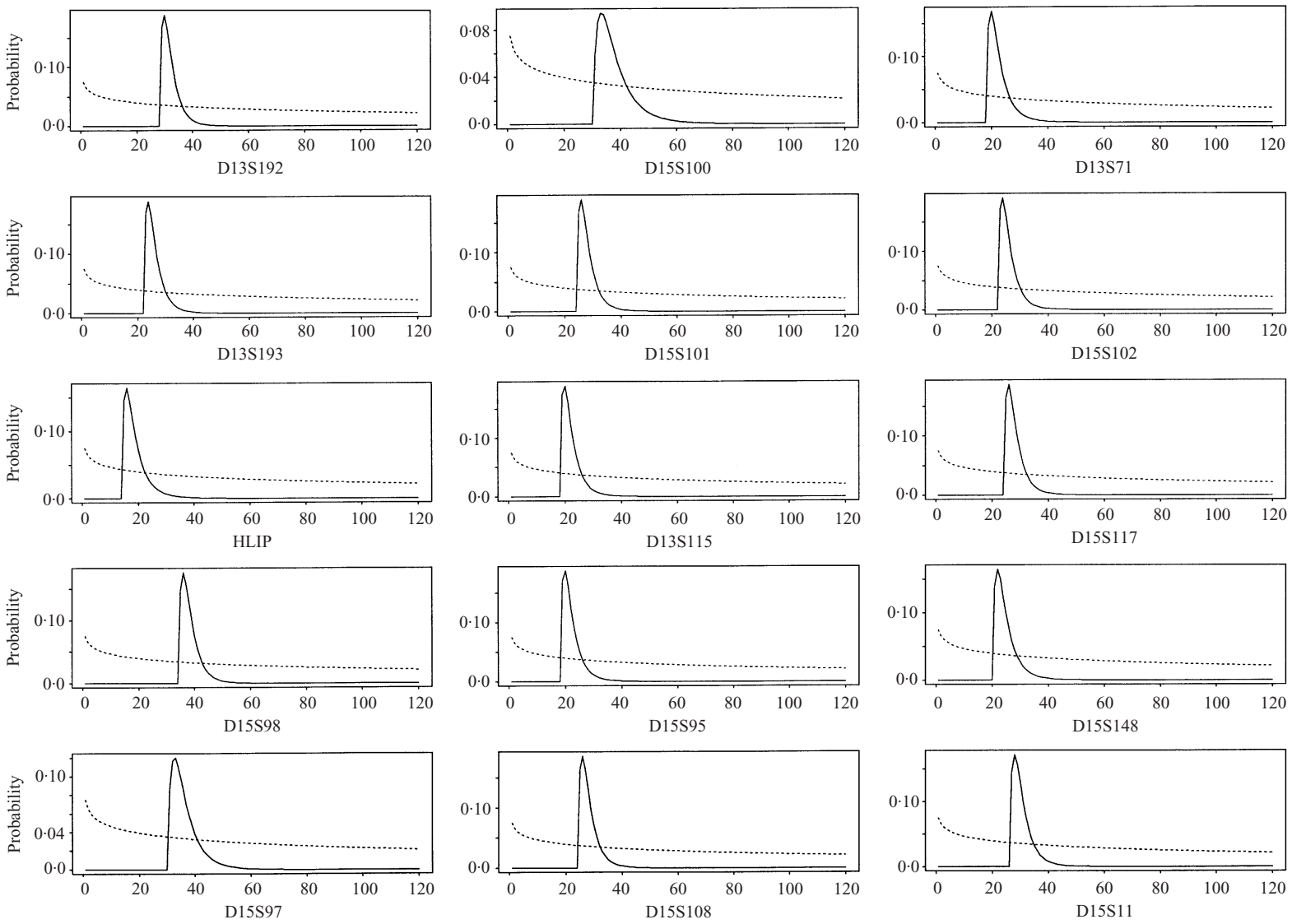

Fig. 4. Posterior distributions of range given the working prior (last 15 loci).

by increasing the probability value of repeat numbers close to zero:

$$p(\rho) = \frac{(1001-\rho)^7}{\sum_{k=0}^{1000}(1001-k)^7} \cdot \mathbf{I}_{\{0,\ldots,1000\}}(\rho). \quad (12)$$

A graphical representation of the prior distributions for ρ is shown in Fig. 1. The scale on which we are able to define boundaries is that of α and β; thus (10) must be considered as the working prior distribution. Equation (11) refers to a prior in which the set $\{0,\ldots,400\}$ contains range values with similar plausibility. The prior in (12) reflects a belief that a range greater than 300 is unlikely, while values smaller than 200 are more plausible. The choice of prior distribution is not straightforward if the amount of prior information is not large.

(iv) *The case study*

The data previously published by Bowcock *et al.* (1994) were re-analysed to estimate range values at 30 microsatellites loci. They include 30 dinucleotide microsatellite loci studied in samples from 14 biological populations (Australian, Cambodian, CAR Pygmy, Chinese, North Italian, Japanese, Karitiana, Lisongo, Melanesian, North European, New Guinean, Surui, Zaïre Pygmy), as shown in Table 1. The original biological populations were fully aggregated into one macropopulation, called the reference population, because α and β are assumed to be a feature of a given locus, and not to differ among the populations.

3. Results and discussion

The analysis of 30 microsatellite loci is performed assuming independence among loci. In Fig. 2, barplots of repeat numbers are shown for each locus. On the x-axis the name of each locus is reported below the repeat scores, while probability values are located on the y-axis. A fair amount of heterogeneity is present among loci, both in the shape of the distribution and in the minimum and maximum repeat numbers, and therefore for the range. The microsatellite locus D13S133 (Fig. 2, second row, fifth column from left) differs greatly from the other loci, in both the pronounced multimodality and the large range.

The visual insight from Fig. 2 is confirmed by further numerical analysis. The three most represented repeat numbers in the sample constitute from 0·47 to 0·90 of the overall frequency distribution (data not shown). The absolute mean difference at a locus among pairs of repeat numbers is typically above 3·0, with only one locus in which its value is 2·22 and one locus in which it reaches a value of 22.

The sampled values of range, minimum, maximum for each locus are listed in Table 1, from left to right. The observed value of the range r varies from 10 to 70 for locus D13S133. A substantial amount of variability is found in the minimum and maximum values of repeat numbers. Note that the sample size n is greater than 240 at each locus.

Table 2. *Test of hypothesis*

me	*mo*	*c*2	0	1	2	3	4	5	6	7
24·0	21	32	14	17	15	12	10	7	6	19
21·4	19	28	17	19	16	12	9	7	5	14
36·5	30	52	5	8	9	9	8	7	7	47
17·4	15	24	18	19	16	12	9	6	5	14
34·9	31	44	10	13	13	12	10	8	7	27
22·0	19	30	14	17	15	12	10	7	6	19
13·6	10	22	19	19	15	11	8	6	5	16
32·5	29	41	11	14	14	12	10	8	6	24
19·3	17	26	18	19	16	12	9	7	5	14
25·7	23	33	15	17	15	13	10	7	6	17
75·1	72	84	9	12	12	12	10	9	7	29
27·3	25	34	17	19	16	13	10	7	5	14
36·8	33	46	10	14	13	12	10	8	7	26
32·9	26	50	5	8	8	8	8	7	7	48
23·3	21	30	17	19	16	13	9	7	5	14
31·2	29	37	17	19	16	13	10	7	5	13
25·3	23	32	17	19	16	13	10	7	5	14
18·3	15	27	15	16	14	12	9	7	6	21
37·5	35	44	15	18	16	13	10	8	6	15
35·6	32	46	9	12	12	11	10	8	7	32
37·6	32	52	6	9	9	9	9	8	7	43
27·3	25	34	17	19	16	13	10	7	5	14
21·3	19	28	17	19	16	12	9	7	5	14
21·4	19	28	17	19	16	12	9	7	5	15
27·3	25	34	16	19	16	13	10	7	5	14
22·0	19	30	15	17	15	12	10	7	6	19
25·3	23	32	17	19	16	13	9	7	5	13
27·3	25	34	17	19	16	13	10	7	5	14
24·0	21	32	14	16	15	12	10	8	6	19
29·7	27	37	14	17	16	13	10	8	6	17

From left to right, the mean (*me*), the mode (*mo*), the right bound of the 95% HPD region (*c*2) and the posterior probability values (times 100) of the multiple test of the eight hypotheses in (5) are shown. The left bound of the 95% HPD region is equal to the plug-in range estimate (Table 1). Hypotheses are labelled as 0 for $H^{(0)}$, 1 for $H^{(1)}$, etc.

The working prior distribution in (10) was used to obtain the posterior distributions of the range at 30 microsatellite (Figs. 3, 4). For all 30 loci, the posterior distributions are highly concentrated with respect to the working prior, as the latter was deliberately chosen to be diffuse.

As regards point estimates (Table 2), range estimates based on the mean values of the posterior distributions for the four priors differ only slightly from the observed range r. Estimates based on the mode are often unchanged or else usually differ by only one repeat unit. Point estimates based on the mode are closer to the observed values. We suggest classifying the microsatellite loci according to the mode of the posterior distribution obtained using (10).

At each locus, the HPD region (Table 2) includes the plug-in estimate of range (Table 1), and its size is

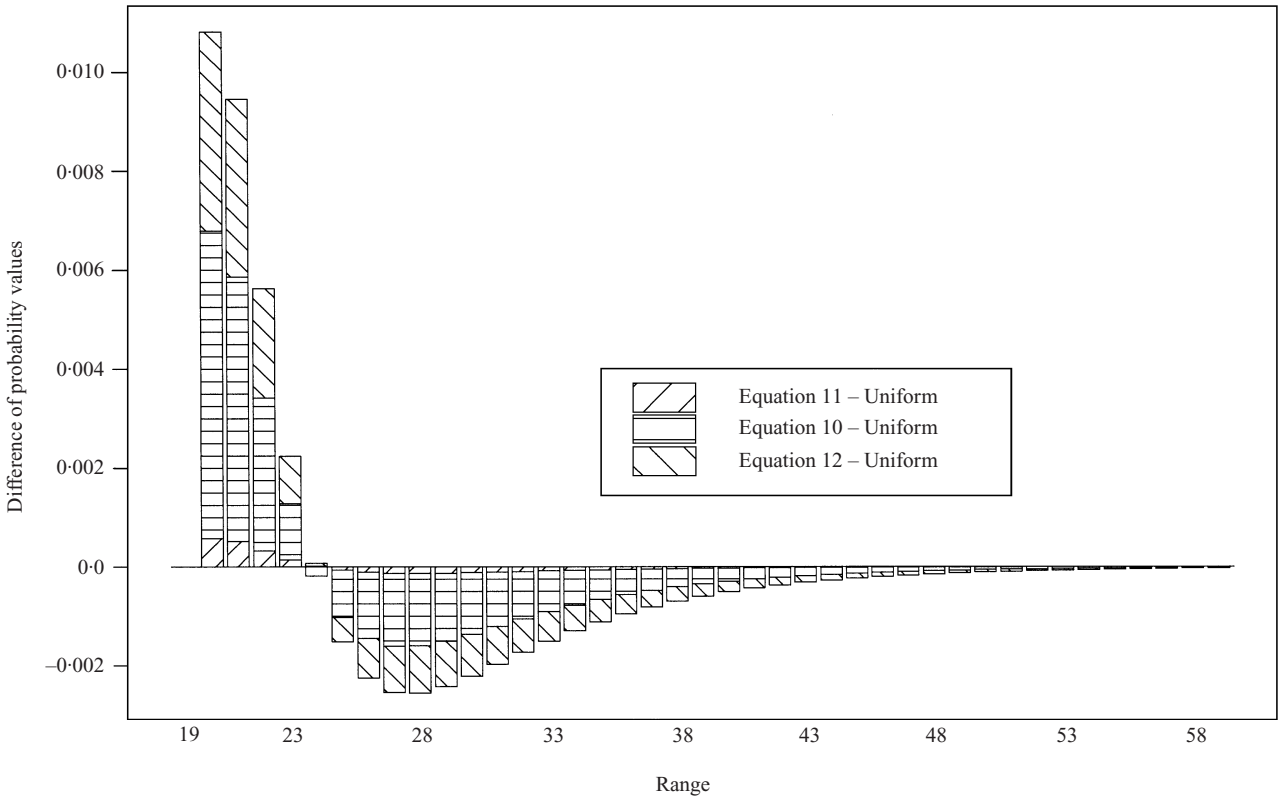

Fig. 5. Differences between probability values of the posterior distributions of ρ at locus D13S270. Probability values from the posterior distribution under a uniform prior for ρ are subtracted from the values of the final distributions of ρ obtained using priors in (10), (11) and (12).

typically equal to 10. The posterior probability values of the hypotheses in (5) are smaller than 50% and are often quite similar. We conclude that on the basis of the overall available information there is not a range value that stands out as most plausible for the population range, because posterior probability values in the multiple test are quite similar. This conclusion is also evident from the size of HPD regions and from the values of Bayes factors (data not shown).

The robustness of the posterior distribution for the range was studied by examining the changes in the posterior due to the use of different priors for ρ. The differences in probability values are difficult to judge in an overlapped plot of posterior distributions at one locus. In Fig. 5, the difference between the probability values from the prior distributions in (10, 11, 12) and values from the uniform prior are represented in a stacked barplot to summarize the overall behaviour of locus D13S270. The differences in probability values are in general quite small for each repeat number.

As regards robustness of the choice of prior distribution for $\underline{\pi}$, we also performed the integration for $\pi_i = 10/(\beta-\alpha+1)$ at each i, and the integrated likelihoods turned out to be more concentrated than if $\pi_i = 1/(\beta-\alpha+1)$ (data not shown). The choice $\lambda = 10$ represents increased strength of our prior belief, and it can be interpreted as a virtual sample of size 10. We did not try greater values because the resulting strength is not justified on the basis of our prior knowledge.

The choice of a working prior with $\pi_i = 1/(\beta-\alpha+1)$ entails that the probability for each repeat at one locus is likely to be very small or very large (U-shaped distributions). This is an inevitable consequence of choosing small values for λ, to allow small samples to affect substantially the shape of the distribution. This behaviour would not occur for higher values, say $\lambda = 1000$.

Readers interested in numerical summaries of the posterior distributions may find the posterior probability values in a file at the web page of one of the authors (http://www.ds.unifi.it/~stefanin).

This Bayesian framework has several advantages for range estimation. As weak as the prior information might seem, the Bayesian approach takes advantage of it in a coherent way (O'Hagan, 1994, cap. 1). That is, we used the information available about the asymptotic distribution of repeat numbers at (genetic) equilibrium (in building the prior distribution of $\underline{\pi}$) and about a reasonable set of values for the boundaries of repeat scores in dinucleotides (in obtaining the prior distribution of α and β). In addition, Bayesian computation can be accomplished here without performing Monte Carlo simulations (e.g. Stefanini & Feldman, 1998), namely with minimum effort. Estimates of the range may change sharply as more

experimental information becomes available, but Bayesian updating of the estimates adequately reflects this.

The model assumes that the sharp bounds α and β exist as features independent of the biological populations, and that alleles carrying repeat numbers equal to α and β belong to the population from which the sample is drawn, i.e. π_α, π_β are greater than zero at each locus. The existence of sharp bounds is still an open question. Recently, Kruglyak *et al.* (1998) proposed a model that explains heterogeneity in the distributions of repeat sizes for different organisms and repeat motifs without imposing sharp bounds. Fu & Chakraborty (1998) developed a simultaneous estimation procedure for all the parameters of the stepwise mutation model based on Monte Carlo samples. An extension of our model to parallel these studies should include information about the populations dynamics under the stepwise mutation model, but this would be unlikely to allow closed-form solutions.

Our analysis does not depend on a specific model for the population dynamics (e.g. Wilson & Balding, 1998) other than the genetic features contained in the assumptions above. This approach is a reasonable alternative to a full Bayesian hierarchical model that might take account of all the different genetic models (stepwise, infinite alleles, etc.), weighted by the corresponding prior belief, because of the prohibitive computational burden, even under Monte Carlo computation. Likewise, the search for an unbiased estimator in a non-Bayesian setting is less appealing than usual, because several models are expected to fit the data quite well. Approximately unbiased estimators of range seem to depend on the adoption of very simple models in which the likelihood function is typically so constrained as to be almost useless.

Further investigations might include information about the expected number of alleles maintained in the population in the prior distribution of our model. Recently, a simultaneous estimation algorithm for the key genetic parameters was proposed by Fu & Chakraborty (1998). Results by Dib *et al.* (1996) and Weber & Wong (1993) suggest that a mutation rate equal to 0·0006 might be appropriate for dinucleotide microsatellites. If the assumption that α and β are represented in the population is not satisfied, then ρ can be interpreted as the minimum reasonable value of $\beta-\alpha$.

Finally, the nuisance parameters in $\underline{\pi}$ play an important role. Indeed α and β depend on $\underline{\pi}$ at least until experiments show that alleles greater (or smaller) than a certain value are not compatible with survival of the organism. The distribution of repeat numbers at a locus changes over time, and neighbouring repeat numbers may have correlated frequencies under a stepwise mutation model (Feldman *et al.*, 1997). A correlation parameter was not considered because the stepwise genetic model does not belong to the conditioning information of our model. We have assumed that the time scale of changes in the distribution is large enough to make the proposed model useful.

We thank E. Minch for providing the dataset and two anonymous referees for comments that improved the manuscript. F.M.S. thanks the Morrison Institute for Population and Resource Studies, at Stanford, where part of this research was performed. This research was supported in part by NIH grants GM 28016 and GM 28428.

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.

Bernardo, J. M. & Smith, A. F. M. (1994). *Bayesian Theory*. New York: Wiley.

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R. & Cavalli-Sforza, L. L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **386**, 455–457.

Casella, G. C. & Berger, R. L. (1990). *Statistical Inference*. Belmont, CA: Duxbury Press.

Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., Lathrop, M., Gyapay, G., Morissette, J. & Weissenbach, J. (1996). A comprehensive genetic map of the human genome based on 5264 microsatellites. *Nature* **380**, 152–154.

Feldman, M. W., Bergman, A., Pollock, D. D. & Goldstein, D. B. (1997). Microsatellite genetic distances with range constraints: analytic description and problems of estimation. *Genetics* **145**, 207–216.

Fu, Y. & Chakraborty, R. (1998). Simultaneous estimation of all the parameters of a stepwise mutation model. *Genetics* **150**, 487–497.

Garza, J. C., Slatkin, M. & Freimer, N. B. (1995). Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Molecular Biology and Evolution* **12**, 594–603.

Goldstein, D. B., Ruiz-Linares, R. A., Cavalli-Sforza, L. L. & Feldman, M. W. (1995). An evaluation of genetic distances for use with microsatellite loci. *Genetics* **139**, 463–471.

Kruglyak, S., Durret, R. T., Schug, M. D. & Aquadro, C. F. (1998). Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proceedings of the National Academy of Sciences of the USA* **95**, 10774–10778.

Moran, P. A. P. (1975). Wandering distributions and the electrophoretic profile. *Theoretical Population Biology* **8**, 318–330.

Nauta, M. J. & Weissing, F. J. (1996). Constraints on allele size at microsatellite loci: implications for genetic differentiation. *Genetics* **143**, 1021–1032.

O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics*, vol. 2b, *Bayesian Inference*. London: Edward Arnold.

Ohta, T. & Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research* **22**, 201–204.

Pollock, D. D., Bergman, A., Feldman, M. W. & Goldstein, D. B. (1998). Microsatellite behavior with range constraints: parameter estimation and improved distance

estimation for use in phylogenetic reconstruction. *Theoretical Population Biology* **53**, 256–271.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457–462.

Stefanini, F. M. (1997). Metodologia Bayesiana gerarchica ed informazione evolutiva in loci genetici microsatellite. PhD thesis, Department of Statistics, University of Florence, Italy.

Stefanini, F. M. & Feldman, M. W. (1998). Microsatellite loci and the origin of modern humans: a Bayesian analysis. In *Evolutionary Theory and Processes: Modern Perspectives* (ed. S. P. Wasser), pp. 249–269. Dordrecht: Kluwer.

Valdes, A. M., Slatkin, M. & Freimer, N. B. (1993). Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* **133**, 737–749.

Weber, J. L. & Wong, C. (1993). Mutation of human short tandem repeats. *Human Molecular Genetics* **2**, 1123–1128.

Wilson, I. J. & Balding, D. J. (1998). Genealogical inference from microsatellite data. *Genetics* **150**, 499–510.

Zhivotovsky, L. A. & Feldman, M. W. (1995). Microsatellite variability and genetic distances. *Proceedings of the National Academy of Sciences of the USA* **92**, 11549–11552.

Zhivotovsky, L. A., Feldman, M. W. & Grishechkin, S. A. (1997). Microsatellite evolution with biased mutations. *Molecular Biology and Evolution* **14**, 926–933.