Published online 29 June 2012

doi:10.1093/nar/gks644

Detection of dispersed short tandem repeats using reversible jump Markov chain Monte Carlo

Tong Liang[1], Xiaodan Fan[2,*], Qiwei Li[2] and Shuo-yen R. Li[1]

[1]Department of Information Engineering and [2]Department of Statistics, Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

Received April 10, 2012; Revised May 23, 2012; Accepted June 11, 2012

ABSTRACT

Tandem repeats occur frequently in biological sequences. They are important for studying genome evolution and human disease. A number of methods have been designed to detect a single tandem repeat in a sliding window. In this article, we focus on the case that an unknown number of tandem repeat segments of the same pattern are dispersively distributed in a sequence. We construct a probabilistic generative model for the tandem repeats, where the sequence pattern is represented by a motif matrix. A Bayesian approach is adopted to compute this model. Markov chain Monte Carlo (MCMC) algorithms are used to explore the posterior distribution as an effort to infer both the motif matrix of tandem repeats and the location of repeat segments. Reversible jump Markov chain Monte Carlo (RJMCMC) algorithms are used to address the transdimensional model selection problem raised by the variable number of repeat segments. Experiments on both synthetic data and real data show that this new approach is powerful in detecting dispersed short tandem repeats. As far as we know, it is the first work to adopt RJMCMC algorithms in the detection of tandem repeats.

INTRODUCTION

A tandem repeat is a stretch of sequence composed of multiple adjacent approximate copies of a particular substring. Tandem repeats have been found abundant in both DNA (1) and protein sequences (2) of most species. They have played critical roles in genome evolution because of frequent recombination or slippage events (3). There is also an increasing amount of researches showing that tandem repeats are related to many human diseases (4), such as Huntington's disease (5) and cancer (6). On the positive side, tandem repeats can also benefit by generating functional variability and allowing swift adaptive evolution of certain traits (7,8). As a powerful tool, tandem repeats are frequently used for genetic mapping (9), genotyping (10) and forensics studies (11).

Tandem repeats differ in the conservation of their pattern. They can be strongly conservative as in $....gt\underbrace{ATCC}\underbrace{ATCC}\underbrace{ATCC}cg....$, where the repeat unit *ATCC* occurs three times consecutively. They can also be divergent as in $....gt\underbrace{ATCA}\underbrace{ATAC}\underbrace{ATCC}cg...$, which allows some mismatches between the repeat units. The two versions are called exact and approximate tandem repeats, respectively. In the remainder of this article, we focus on the latter case, which is more general and harder to detect. There are mainly two ways to represent a sequence pattern (i.e. a motif): motif consensus and motif matrix (12). Using the previous approximate tandem repeat as an example, the pattern can be represented by the motif consensus as '*ATCC*' or by the motif matrix

$$\begin{array}{c c} & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{c} A \\ T \\ C \\ G \end{array} & \begin{pmatrix} 1 & 0 & 1/3 & 1/3 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2/3 & 2/3 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array},$$

where each column of the matrix denotes the relative frequencies of the four nucleotides (A, T, C, G) showing at the corresponding position.

Over the past decade, a number of softwares have been developed to detect approximate tandem repeats. Generally speaking, current repeat detection methods can be roughly classified as either a string matching approach or a signal processing approach. The string matching approach detects repeats by scoring the sequence alignment between the input sequence and a library of curated repeat consensus, *k*-mers (a segment composed of *k* nucleotides or amino acids) or itself. RepeatMasker (Smit, AFA, Hubley, R and Green, P. RepeatMasker Open-3.0.), RECON (13), REPuter (14), mreps (15), Tallymer (16) and TRF (17) are some representative softwares in this class. The signal processing approach is mainly based on periodicity in the sequence.

*To whom correspondence should be addressed. Tel: +852 3943 7930; Fax: +852 2603 5188; Email: xfan@sta.cuhk.edu.hk

It uses techniques such as discrete Fourier transform, short-time periodicity transform, exact periodic subspace decomposition and autoregressive modeling to perform spectral analysis (18–21). A more comprehensive review and a performance comparison of existing softwares can be found in (22,23).

A consecutive block composed of repeat units is called a tandem repeat segment. Multiple tandem repeat segments sharing a same pattern may be dispersively distributed in a sequence. Current methods are largely based on the motif consensus representation and a sliding-window technique. Our work concentrates on the situation where multiple tandem repeat segments composed of instances of the same short motif pattern are dispersively distributed in a sequence. We name the repeats in this scenario as the Dispersed Short Approximate Tandem Repeats (DSATRs). In this article, the pattern of DSATR is described by a motif matrix and will be inferred through a Bayesian approach. Markov chain Monte Carlo (MCMC) algorithms are used to explore the complex posterior distribution of interested parameters. Bayesian inference by iterative sampling has been hastily developed in the past decades (24,25). It has been used to detect the binding sites in DNA and protein sequences, such as the Gibbs Motif Sampler (GMS) (26,27). Other popular algorithms for binding motif detection include AlignACE (28) and MEME (29). Jensen *et al.* (30) provided a review of algorithms for binding motifs. In the binding sites problem, the motif instances are dispersedly distributed in multiple input sequences or multiple dispersed locations in a sequence. Although for tandem repeats, the motif instances (i.e. repeat units) are also locally grouped together. Based on Gibbs sampler, Li *et al.* (31) dealt with the case where each of the multiple input sequences contains a gapped repeat segments of the same repeat pattern.

Different from classical motif discovery problems, the unknown number of repeat segments in the target sequence leads to a transdimensional model selection problem. One possible solution for this problem is to use the reversible jump Markov chain Monte Carlo (RJMCMC) algorithm introduced by Green (32). Brooks *et al.* (33) investigated on the efficient construction of RJMCMC algorithms. Al-Awadhi *et al.* (34) also discussed how to improve the acceptance rate of RJMCMC. However, although RJMCMC is theoretically capable in solving transdimensional problems, its slow convergence limited its application to simple model selection problems. Successful applications of RJMCMC in real problems are rarely seen in the literature. As an alternative solution for transdimensional model selection, the birth-and-death approach is introduced in (35). A comparison between RJMCMC and the birth-and-death approach is provided by Cappe *et al.* (36). A geometric approach is presented in (37) for transdimensional MCMC. The method of (38), relying on the asymptotic behaviors of different risk functions and abstract semi-parametric bootstrap principles, targeted at the model selection problem for variable length Markov chain.

In this article, a Bayesian approach is developed to detect DSATR in a *de novo* fashion. We detect the motif matrix and locate the motif instances simultaneously through their joint posterior distribution, which assesses the fitness of all points in the parameter space to the data in a probabilistic view. RJMCMC algorithms are adapted to tackle the problem that the number of segments dispersed on the whole sequence is unknown. Extra effort is devoted to speeding up the convergence of RJMCMC.

MATERIALS AND METHODS

A generative model for a sequence with DSATR

A schematic view of a sequence with DSATR is shown in Figure 1. The input sequence is denoted as $\boldsymbol{R} = (r_1, r_2, \ldots, r_L)$, where the nucleotide at the l-th location, r_l, takes a value from the alphabet {A,T,C,G}, and L denotes the length of the sequence. The input sequence consists of two parts, namely the tandem repeat region and the background region (i.e. the non-tandem repeat region). The tandem repeat region is composed of one or multiple separated repeat segments. We define a repeat segment with k repeat units as k adjacent instances of the same motif, i.e. a contiguous block with kw nucleotides in the sequence, where w denotes the width of the motif. Notice that the number of repeat segments is unknown in advance. Given the sequence $\boldsymbol{R}$, our goal is to infer the motif matrix of the tandem repeats and the locations of repeat segments. We model every nucleotide r_l in the sequence as a multinomial random variable. More specifically, for the nucleotides inside the repeat unit, we model them using a product multinomial (PM) distribution parameterized by a 4-by-w matrix $\boldsymbol{\Theta}$ (27). We call the repeat pattern, represented by $\boldsymbol{\Theta}$, the sequence motif of the tandem repeats, where $\boldsymbol{\Theta} = [\theta_1, \theta_2, \ldots, \theta_w]$ and $\theta_i = (\theta_{i,1}, \theta_{i,2}, \theta_{i,3}, \theta_{i,4})^{\mathrm{T}}$ representing the proportion of the four nucleotides {A,T,C,G} at the i-th position of the repeat unit. These parameters satisfy that $\sum_j \theta_{i,j} = 1$ and $\theta_{i,j} \geq 0$ for all i, j. The nucleotides in the background region are modeled as independent samples from the multinomial distribution parameterized by θ_0, where $\theta_0 = (\theta_{0,1}, \theta_{0,2}, \theta_{0,3}, \theta_{0,4})^{\mathrm{T}}$ representing the proportion of the four nucleotides {A,T,C,G} at the positions outside the repeat region.

As shown in Figure 1, we assume that the input sequence is generated in three steps. We first use θ_0 to generate a background sequence, i.e. the non-repeat region with length L_{b}. In the second step, we sample the number of repeat segments G from a random distribution and then construct the G repeat segments. To model the repeat segments, we introduce a vector $\boldsymbol{K} = (k_1, k_2, \ldots, k_G)^{\mathrm{T}}$, where k_g is a random variable indicating the number of repeat units in the g-th segment. Each repeat unit in the repeat segments is independently sampled from the PM distribution parameterized by $\boldsymbol{\Theta}$. Finally, we randomly choose G different positions from the background sequence and insert the G repeat segments into it. We denote the final locations of the G repeat segments in the final sequence $\boldsymbol{R}$ of length $L = L_{\mathrm{b}} + \sum_g k_g w$ as $\boldsymbol{A} = (a_1, a_2, \ldots, a_G)^{\mathrm{T}}$, where $1 \leq a_g \leq L - w + 1$.

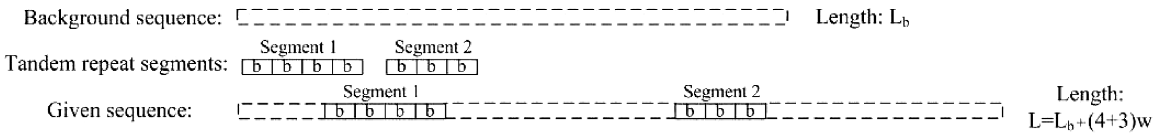

Figure 1. A schematic view of a sequence with DSATR. A background sequence and multiple repeat segments are generated independently. All repeat units in these repeat segments are random instances of a common motif 'b' of width w. The input sequence is generated by randomly inserting the repeat segments into separated locations in the background sequence.

This generating procedure avoids the overlapping and adjacency between different repeat segments. The output of this three-step procedure is a sequence $\boldsymbol{R}$ of length L, which includes G tandem repeat segments composed of instances of the same motif matrix $\mathbf{\Theta}$.

Statistical inference

For the input sequence $\boldsymbol{R}$, which is the only observed data, we are interested in estimating all parameters $\{\theta_0, \mathbf{\Theta}, \boldsymbol{A}, \boldsymbol{K}, G\}$. At this stage, we assume that the width of repeat unit w is given in advance. We use a Bayesian method to detect the tandem repeat segments modeled in the previous section. The joint posterior distribution of all parameters $\pi(\theta_0, \mathbf{\Theta}, \boldsymbol{A}, \boldsymbol{K}, G|\boldsymbol{R})$ will be used to make statistical inference.

Likelihood and prior distributions

For each tandem repeat segment, the starting position and the number of repeat units are used to characterize it. We denote the g-th segment as $\boldsymbol{S}(a_g, k_g)$, where $\boldsymbol{S}(a_g, k_g) = \{r_{a_g+j} : 0 \leq j \leq k_g w - 1\}$. Let $\boldsymbol{R}_{TR(i)}$ denote the set of the i-th nucleotides of all repeat units and $\boldsymbol{R}_{\overline{TR}}$ denote the set of nucleotides in the non-repeat region. For convenience, we define the following indicator function I: $I(r,*) = 1$ if r is equal to $*$; otherwise, $I(r,*) = 0$. Here * takes a value from {A,T,C,G}. We introduce a counting function $\boldsymbol{H}$ such that $\boldsymbol{H}(\boldsymbol{F}) = (h_{A,\boldsymbol{F}}, h_{T,\boldsymbol{F}}, h_{C,\boldsymbol{F}}, h_{G,\boldsymbol{F}})^{\mathrm{T}}$, where $\boldsymbol{F}$ is any set of nucleotides and $h_{*,\boldsymbol{F}} = \sum_{r \in \boldsymbol{F}} I(r, *)$. For any vectors $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^{\mathrm{T}}$ and $\mathbf{v} = (v_1, \ldots, v_d)^{\mathrm{T}}$, we define that $\boldsymbol{\mu} + \mathbf{v} = (\mu_1 + v_1, \ldots, \mu_d + v_d)^{\mathrm{T}}$, $\boldsymbol{\mu}^{\mathbf{v}} = ({\mu_1}^{v_1}, \ldots, {\mu_d}^{v_d})^{\mathrm{T}}$, and $|\mu^{\nu}| = \prod_{i=1}^{d} {\mu_i}^{\nu_i}$. With these notations, the complete likelihood can be factorized into a product of probabilities for the repeat region and the non-repeat region as follows:

$$\pi(\boldsymbol{R}|\theta_0, \mathbf{\Theta}, \boldsymbol{A}, \boldsymbol{K}, G) = |\theta_0^{\boldsymbol{H}(\boldsymbol{R}_{\overline{TR}})}| \prod_{i=1}^{w} |\theta_i^{\boldsymbol{H}(\boldsymbol{R}_{TR(i)})}|.$$

Let $p(\cdot)$ denote the prior distribution of a parameter. We specify the prior for $\mathbf{\Theta}$ and θ_0 independent of other parameters. Thus, we can decompose the prior for $\{\theta_0, \mathbf{\Theta}, \boldsymbol{A}, \boldsymbol{K}, G\}$ as

$$p(\theta_0, \mathbf{\Theta}, \boldsymbol{A}, \boldsymbol{K}, G) = p(\theta_0) p(\mathbf{\Theta}) p(G) p(\boldsymbol{K}|G) p(\boldsymbol{A}|\boldsymbol{K}, G).$$

For computational convenience, we use conjugate priors for θ_i, i.e. a Dirichlet (Dir) distribution with parameter β_i. We denote $\mathbf{B} = [\beta_1, \beta_2, \ldots, \beta_w]$, where each θ_i is associated with an 4-by-1 vector β_i. The parameter $\mathbf{\Theta}$ therefore follows a product Dirichlet (PD) distribution parameterized by the matrix $\mathbf{B}$, i.e. $\mathbf{\Theta} \sim$ PD ($\mathbf{B}$). Similarly we assume $\theta_0 \sim$ Dir (β_0). We further assume G takes integer values from $G_{\min}$ to $G_{\max}$ with equal probability. That is, G follows a discrete uniform (DU) distribution, which is denoted as $G \sim$ DU $[G_{\min}, G_{\max}]$. Similarly we assume $k_g \sim$ DU $[k_{\min}, k_{\max}]$.

To avoid the overlap and adjacency between the different tandem repeat segments and guarantee the identifiability, we require that $\max\{1, a_{g-1} + k_{g-1} w + 1\} \leq a_g \leq \min\{a_{g+1} - k_g w - 1, L - k_G w + 1\}$. In practice, there is little prior knowledge about $\boldsymbol{A}$. Given $\boldsymbol{K}$ and G, we assume that $\boldsymbol{A}$ is evenly distributed on all possible choices. Thus we have $p(\boldsymbol{A}|\boldsymbol{K}, G) = \binom{L - \sum_{g=1}^{G} k_g w + 1}{G}^{-1}$. Given G, we assume that the elements of $\boldsymbol{K}$ are mutually independent, i.e. $p(\boldsymbol{K}|G) = \prod_{g=1}^{G} p(k_g)$, where $p(k_g) = 1/(k_{\max} - k_{\min} + 1)$. Thus, we get the prior for all interested parameters as

$$p(\theta_0, \mathbf{\Theta}, \boldsymbol{A}, \boldsymbol{K}, G) = p(\theta_0) p(\mathbf{\Theta}) p(G) \binom{L - \sum_{g=1}^{G} k_g w + 1}{G}^{-1} \prod_{g=1}^{G} p(k_g).$$

The formula implies that the prior $p(\theta_0, \mathbf{\Theta}, \boldsymbol{A}, \boldsymbol{K}, G)$ varies with θ_0, $\mathbf{\Theta}$, k_g and G, but not with the specific values in $\boldsymbol{A}$. We can now write down the full posterior probability as

$$\pi(\theta_0, \mathbf{\Theta}, \boldsymbol{A}, \boldsymbol{K}, G|\boldsymbol{R}) \propto p(\theta_0, \mathbf{\Theta}, \boldsymbol{A}, \boldsymbol{K}, G) \pi(\boldsymbol{R}|\theta_0, \mathbf{\Theta}, \boldsymbol{A}, \boldsymbol{K}, G)$$
$$= \frac{\prod_{i=0}^{w} \mathrm{Dir}(\beta_i) \Gamma(G+1) \Gamma(L - \sum_{g=1}^{G} k_g w + 2 - G)}{(G_{\max} - G_{\min} + 1) \Gamma(L - \sum_{g=1}^{G} k_g w + 2)} \times \frac{|\theta_0^{\boldsymbol{H}(\boldsymbol{R}_{\overline{TR}})}| \prod_{i=1}^{w} |\theta_i^{\boldsymbol{H}(\boldsymbol{R}_{TR(i)})}|}{(k_{\max} - k_{\min} + 1)^G}. \quad \mathbf{(1)}$$

Sampling from posterior distribution

Our MCMC algorithm is composed of three steps. We first use a Gibbs sampling algorithm to explore the posterior distribution when G is given. Then, we use RJMCMC to update G. Two versions of RJMCMC are designed and compared. Finally, we will discuss extra moves to further improve the efficiency of the sampling algorithm.

Gibbs sampling when G *is given*. The joint posterior distribution of θ_0, $\mathbf{\Theta}$, $\boldsymbol{A}$ and $\boldsymbol{K}$, while $\boldsymbol{R}$ and G are given, can be explored by a Gibbs sampling algorithm. We iteratively sample all our interested parameters except G via corresponding conditional posterior probability as follows.

We can obtain the conditional posterior distribution of $\boldsymbol{\Theta}$ from the full posterior probability as follows

$$\pi(\boldsymbol{\Theta}|\boldsymbol{R}, \theta_0, \boldsymbol{K}, \boldsymbol{A}, G) \propto \prod_{i=1}^{w} |\theta_i^{\boldsymbol{H}(\boldsymbol{R}_{TR(i)})+\beta_i}|. \tag{2}$$

Here a conjugate prior $\boldsymbol{\Theta} \sim \text{PD}$ ($\mathbf{B}$) leads to the conditional posterior distribution $\boldsymbol{\Theta}|\boldsymbol{R}, \theta_0, \boldsymbol{K}, \boldsymbol{A}, G \sim \text{PD}$ ($\boldsymbol{Q}+\mathbf{B}$), where $\boldsymbol{Q} = [\boldsymbol{H}(\boldsymbol{R}_{TR(1)}), \boldsymbol{H}(\boldsymbol{R}_{TR(2)}), \ldots, \boldsymbol{H}(\boldsymbol{R}_{TR(w)})]$. Similarly, we derive the conditional posterior distribution of θ_0

$$\pi(\theta_0|\boldsymbol{R}, \boldsymbol{\Theta}, \boldsymbol{A}, \boldsymbol{K}, G) \propto |\theta_0^{\boldsymbol{H}(\boldsymbol{R}_{\overline{TR}})+\beta_0}|. \tag{3}$$

Let $\boldsymbol{A}_{[-g]}$ and $\boldsymbol{K}_{[-g]}$ denote all elements of $\boldsymbol{A}$ excluding a_g and all elements of $\boldsymbol{K}$ excluding k_g, respectively. As pointed out previously, the prior $p(\theta_0,\boldsymbol{\Theta}, \boldsymbol{A}, \boldsymbol{K}, G)$ keeps invariant as we update a_g. Thus, we obtain the following conditional posterior distribution for a_g

$$\begin{aligned}&\pi(a_g|\boldsymbol{R}, \theta_0, \boldsymbol{\Theta}, \boldsymbol{K}, \boldsymbol{A}_{[-g]}, G)\\ &= \frac{\pi(\boldsymbol{R}|\theta_0, \boldsymbol{\Theta}, \boldsymbol{A}_{[-g]}, a_g, \boldsymbol{K}, G)}{\sum_{b=\max\{1,a_{g-1}+k_{g-1}w+1\}}^{\min\{a_{g+1}-k_g w-1,L-k_G w+1\}} \pi(\boldsymbol{R}|\theta_0, \boldsymbol{\Theta}, \boldsymbol{A}_{[-g]}, b, \boldsymbol{K}, G)},\end{aligned} \tag{4}$$

where $\max\{1, a_{g-1} + k_{g-1}w + 1\} \leq a_g \leq \min\{a_{g+1} - k_g w - 1, L - k_G w + 1\}$. It is important to note that the conditional prior $p(\boldsymbol{A}|\boldsymbol{K}, G)$ may change as we update k_g. Thus we obtain the conditional posterior distribution for k_g

$$\begin{aligned}&\pi(k_g|\boldsymbol{R}, \theta_0, \boldsymbol{\Theta}, \boldsymbol{K}_{[-g]}, \boldsymbol{A}, G)\\ &\propto \pi(\boldsymbol{R}|\theta_0, \boldsymbol{\Theta}, \boldsymbol{K}_{[-g]}, k_g, \boldsymbol{A}, G)p(\boldsymbol{A}|\boldsymbol{K}, G),\end{aligned} \tag{5}$$

where k_g takes value from $[k_{\min}, k_{\max}]$.

RJMCMC for updating G. In contrast to other parameters, the updating procedure for G will give rise to the change of the dimensions of $\boldsymbol{A}$ and $\boldsymbol{K}$, which leads to a Bayesian model selection problem. Let $\boldsymbol{M}_G = \{\boldsymbol{A}, \boldsymbol{K}\}$ indicate that both $\boldsymbol{A}$ and $\boldsymbol{K}$ contain G elements. We present a RJMCMC algorithm which aims to sample G and $\boldsymbol{M}_G$ from the conditional posterior distribution $\pi(G,\boldsymbol{M}_G|\theta_0,\boldsymbol{\Theta}, \boldsymbol{R})$. To impose the dimension matching, auxiliary random variables are introduced to propose new repeat segments in the jumping process. The main difficulty in RJMCMC algorithm is how to effectively propose the new segments, i.e. auxiliary variables. Two different versions, namely vanilla and piloted, of the proposal distribution for auxiliary variables are proposed in Section 1 of Supplementary Materials. Since the vanilla RJMCMC has the drawback of slow convergence, we introduce a so-called piloted RJMCMC by constructing the transdimensional move based on one-step-ahead prediction. More specifically, the piloted version takes advantage of the relationship between the input sequence and the current motif model $\boldsymbol{\Theta}$ to propose auxiliary variables. We conclude theoretically that the acceptance rate (i.e. the expected acceptance probability) of the piloted version is larger than that of the vanilla version. Since the major hindrance of RJMCMC from wide applications is its rather slow move around the state space, the increased acceptance rate will then generally improve the convergence rate of the RJMCMC chain. An experimental verification of this point will be given later. The details of RJMCMC are given in Section 1 of Supplementary Materials.

The Gibbs sampling procedure for updating (θ_0,$\boldsymbol{\Theta}$, $\boldsymbol{A}$, $\boldsymbol{K}$) and the RJMCMC step for updating G compose the basic MCMC algorithm for computing our DSATR model. Many factors can affect the efficiency of a MCMC algorithm, such as the high correlation between parameters. We design three extra MCMC moves, namely a local group move, a global group move and a phase-shift move, to improve the mixing of the Markov chain. The details of these moves are given in Section 2 of the Supplementary Materials. A summary of the complete algorithm is provided in Section 3 of the Supplementary Materials. It clearly listed all inputs, outputs, tuning parameters and the detailed procedure of the algorithm.

RESULTS

We have implemented two versions of our complete algorithm in Matlab. The only difference between these two versions is the RJMCMC step, namely the vanilla and piloted versions. To evaluate the model and the algorithm, we apply the proposed algorithm to both synthetic and real data. We will discuss our experiment results and evaluate the performance of our algorithm in terms of convergence and accuracy.

Evaluation and comparison of the two RJMCMC versions using synthetic data

The synthetic sequence with DSATR, which is generated according to the generative model introduced in previous section, is used to compare the performance of two versions in Section 4 of Supplementary Materials. In summary, the piloted version outperforms the vanilla version in terms of convergence speed, the acceptance rate for the RJMCMC moves and the effective sample size within the same number of iterations. But for the accuracy of the statistical inference, they show a similar performance after a sufficient number of iterations.

Comparison with existing methods using synthetic data

Some existing methods can be used to detect DSATR although not specifically designed for this purpose. For a comparison with our algorithm, we tested three widely used methods using synthetic data, including TRF, GMS and RepeatMasker. As non-probabilistic algorithms, both TRF and RepeatMasker construct an explicit alignment score to evaluate a sequence segment and use the score to decide whether to report it as a repeat segment. We explain their scoring functions in Section 4.2.1 of Supplementary Materials.

Synthetic data from the generative model

We begin by defining the signal strength of repeat segments. Here, the signal strength is measured by the degree of conservation of the motif matrix $\boldsymbol{\Theta}$. According

Table 1. Performance comparison on synthetic data sets from the generative model

	w	Conservation	GMS (Std.)	RepeatMasker[a] (Std.)	TRF (Std.)	Vanilla (Std.)	Piloted (Std.)
Sensitivity	3	High	0.258(0.379)	0.966(0.041)	0.967(0.061)	0.976(0.028)	0.980(0.028)
Specificity	3	High	0.597(0.048)[b]	0.931(0.046)	0.896(0.080)	0.967(0.025)	0.969(0.022)
Sensitivity	3	Median	0.000(0.000)	0.724(0.145)	0.662(0.175)	0.918(0.059)	0.920(0.062)
Specificity	3	Median	–[c]	0.936(0.047)	0.932(0.058)	0.937(0.040)	0.944(0.033)
Sensitivity	3	Low	0.000(0.000)	0.232(0.179)	0.241(0.173)	0.771(0.122)	0.782(0.125)
Specificity	3	Low	–[c]	0.952(0.064)[d]	0.908(0.141)[e]	0.907(0.061)	0.905(0.061)
Sensitivity	6	High	0.922(0.026)	0.993(0.010)	0.993(0.009)	0.959(0.073)	0.991(0.010)
Specificity	6	High	0.848(0.037)	0.960(0.024)	0.837(0.104)	0.982(0.014)	0.984(0.014)
Sensitivity	6	Median	0.634(0.049)	0.942(0.051)	0.871(0.100)	0.959(0.046)	0.976(0.020)
Specificity	6	Median	0.790(0.050)	0.965(0.025)	0.906(0.074)	0.976(0.019)	0.977(0.019)
Sensitivity	6	Low	0.192(0.106)	0.310(0.176)	0.216(0.125)	0.901(0.063)	0.909(0.054)
Specificity	6	Low	0.603(0.114)[f]	0.977(0.034)[e]	0.900(0.121)[g]	0.948(0.033)	0.952(0.033)
Sensitivity	9	High	0.989(0.010)	0.976(0.015)	0.998(0.002)	0.953(0.069)	0.988(0.011)
Specificity	9	High	0.959(0.024)	0.983(0.012)	0.763(0.126)	0.988(0.010)	0.989(0.010)
Sensitivity	9	Median	0.804(0.039)	0.954(0.026)	0.937(0.080)	0.944(0.079)	0.981(0.023)
Specificity	9	Median	0.883(0.047)	0.981(0.017)	0.882(0.078)	0.984(0.016)	0.984(0.017)
Sensitivity	9	Low	0.399(0.049)	0.456(0.151)	0.302(0.164)	0.944(0.045)	0.944(0.037)
Specificity	9	Low	0.738(0.077)	0.984(0.021)	0.961(0.057)[h]	0.963(0.022)	0.965(0.024)

Note: Each cell of the table shows the corresponding mean and standard deviation (in the bracket) of the sensitivity or specificity calculated from the 100 different synthetic sequences.

[a] The true consensus is given as the sole repeat pattern in RepeatMasker library, so this comparison favors RepeatMasker.

[b] 68 trails did not report any repeat element.

[c] All trails did not report any repeat element.

[d] 18 trails did not report any repeat element.

[e] 4 trails did not report any repeat element.

[f] 17 trails did not report any repeat element.

[g] 5 trails did not report any repeat element.

[h] 1 trail did not report any repeat element.

to the probability of the dominant nucleotide (39), we define the degree of conservation as high, median or low. Nine data sets of inputs data are generated according to different width (3,6,9) and different signal strength (High, Median and Low) to compare these methods more comprehensively. Here, each data set contains 100 independent synthetic sequences with 5000 nucleotides.

We run our algorithm and the three other programs once on each of the 900 synthetic sequences. To favor the three other programs, we set them as follows. For RepeatMasker, since it needs a repeat library in order to detect corresponding repeats, the dominant nucleotides in each column of motif matrix are selected to build the sole consensus pattern for its repeat library. In other words, RepeatMasker is favored by knowing the true consensus pattern in the input data. This guarantees the high specificity for RepeatMasker. Meanwhile, we set the minimum report score for TRF as 20 (the default value is 50) and the cutoff of RepeatMasker as 100 (the default value is 225) to increase their sensitivities. To favor these three existing methods, we compare the results from all the methods at the nucleotide level. Each nucleotide in the given sequence is labeled as either in repeat region or in background region. The sensitivity is defined as the proportion of true repeat-region nucleotides that are correctly identified, and the specificity is defined as the proportion of the reported repeat-region nucleotides that are true repeat-region nucleotides. More details for the setting of GMS, RepeatMasker and TRF are listed in Section 4 of the Supplementary Materials. In addition, the detailed scoring functions for RepeatMasker and TRF and an experiment for a single sequence with detailed discussion on the sequences are also presented in Section 4 of the Supplementary Materials. For both versions of our RJMCMC algorithm, we run them separately for 3000 iterations with the same prior setting and tuning parameters.

The results are reported in Table 1. It summarizes the sensitivity and specificity of these algorithms on the nine different motif matrixes. Each cell of the table shows the corresponding mean and standard deviation (in the bracket) of the sensitivity or specificity calculated from the 100 different synthetic sequences. In some cases, GMS, RepeatMasker and TRF did not report any repeat elements, therefore the corresponding specificities are not defined.

Table 1 shows that all our algorithms have better performance than GMS for all signal strength and motif width. Since GMS does not make use of the local enrichment of repeat units, it often misses some repeat units of a repeat segment, or even reports only one repeat unit for one repeat segment. Also, the performance of GMS is heavily dependent on the motif width and conservation level. Although our algorithms are able to offset the short width and low conservation of the motif matrix by profiting from the local clustering effect of repeat units.

Although RepeatMasker and TRF are favored by knowing true repeat consensus sequences or high-sensitivity setting, our algorithms generally outperform TRF and RepeatMasker by either better sensitivity at similar specificity, or both better sensitivity and specificity. The advantage of our algorithms is more obvious for short

Table 2. Pairwise comparison of RepeatMasker, TRF and our algorithm on real data

Data and consensus pattern	Algorithm A	Algorithm B (reference) (%)				The percentage of tested sequence classified as repeat region (%)
		Piloted version		TRF	RepeatMasker	
		V1[a]	V2[b]			
Chimp (panTro2)	V1	–	90.41	89.43	90.09	1.32
ChrY 1120001-	V2	100	–	94.27	99.55	1.46
1140000	TRF	76.89	73.29	–	79.73	1.14
(CA)$_n$	RepeatMasker	75.76	75.68	77.97	–	1.11
Dog (Broad/canFan2)	V1	–	93.22	91.93	93.69	57.60
Chr1 3160001-	V2	100	–	98.70	99.33	61.79
3170000	TRF	98.49	98.58	–	98.54	61.71
(CGAAT)$_n$	RepeatMasker	94.57	93.46	92.84	–	58.14
Human (hg19)	V1	–	93.33	92.68	89.47	1.26
Chr2 201650001-	V2	100	–	97.56	92.11	1.35
201670000	TRF	30.16	29.63	–	19.30	0.41
(CA)$_n$	RepeatMasker	80.95	77.78	53.66	–	1.14
Rat (rn3)	V1	–	100	99.20	95.51	2.35
Chr17 24340001-	V2	100	–	99.20	95.51	2.35
24350000	TRF	52.77	52.77	–	51.02	1.25
(TCCTA)$_n$	RepeatMasker	99.57	99.57	100	–	2.45
Zebrafish (danRer6)	V1	–	93.88	92.64	89.89	2.05
Chr13 24220001-	V2	100	–	95.84	96.02	2.18
24250000	TRF	94.30	91.59	–	92.50	2.08
(TA)$_n$	RepeatMasker	95.60	95.87	96.64	–	2.18

[a] V1: Piloted version.
[b] V2: Piloted version with post-processing.

and divergent repeat segments. This is because both TRF and RepeatMasker, as window-based local search methods, cannot integrate the signal from multiple repeat segments. Therefore, the weak signal in individual repeat segments, when treated mutually independently, is likely to be missed by window-based methods. Another drawback of TRF is its incapability in handling the phase-shift between different repeat segments. In addition, we present a synthetic sequence experiment with detailed repeat units to illustrate that the gained efficiency of our algorithm is due to the more efficient modeling and computing, rather than due to different repeats definitions in algorithms, in Section 4.2.2 of Supplementary Materials.

Synthetic data from the coalescent model

To evaluate our algorithm on more realistic data instead of the data purely generated from our generative model, we also used a coalescent model (40–42) to produce a second version of the synthetic data for testing. The details are given in Section 5 of Supplementary Materials. The coalescent model considers the relations between different repeat segments and the mutation for short tandem repeats at both repeat unit and nucleotide levels. The results (see Table S2 in the Supplementary Materials) indicate that our algorithm still outperforms other methods on the data from the coalescent model.

Real data experiment

A real DNA sequence may contain more than one repeat pattern, which will result in multiple local modes in the posterior distribution. Thus we run multiple independent chain from random initial parameter values for real data and report the overall maximum a posteriori (MAP) estimate.

Another concern about the real data is the existence of indels, which has not been directly handled in our algorithm so far because of the lack of experimental knowledge support and technical difficulty. However, our current algorithm can accommodate indels to certain extent due to two reasons. On one hand, since the indels rate in repeat segments is much lower than the substitution rate, the main repeat region and motif matrix can be identified fairly well by our current algorithm without handling indels. On the other hand, if a repeat unit with indels exists in the middle of a repeat segment, our current algorithm can simply treat this repeat unit as background nucleotides and detect the remaining part as two repeat segments. But in case the user prefers reporting the repeat units with short indels, just like what TRF and RepeatMasker do using a special indel step, we also designed a post-processing step to glean repeat units with short indels. The details of the post-processing step are given in Section 6 of Supplementary Materials.

For comparison, a list of real genomic segments from different species are sampled from the Pre-Masked Genomes in RepeatMasker website (http://www.repeatmasker.org/cgi-bin/AnnotationRequest, 21 June 2012, date last accessed).

The detailed settings of the programs are given in Section 6 of Supplementary Materials. The results are summarized in Table 2. Since no ground truth about the repeat pattern and locations in the real data are known, the sensitivity and specificity are not well defined here. Thus this experiment can only provide pairwise coverage comparisons instead of a systematic performance comparison. Each cell in Table 2 represents the proportion of repeat segments identified by the corresponding column algorithm (Algorithm B) that are also identified by the corresponding row algorithm (Algorithm A). Thus, the paired cells (*i*,*j*) and (*j*,*i*) can demonstrate the consistency between the *i*-th and *j*-th algorithms.

The results showed that the post-processing step expanded the reported repeat regions and therefore increased the consistency between our algorithms and other methods. But generally there is a slight inconsistency among the results of different algorithms in most of examples because of the difference in the scoring functions of all three algorithms. RepeatMasker and TRF may differ significantly, e.g. for Human Chr2 201650001-201670000. Our algorithm agreed well with either RepeatMasker or TRF on all cases. The missed repeat units in the results of our algorithm as compared with other methods are mostly due to the occurrence of indels. This is an empirical proof that the lack of indel treatment in our generative model does not seriously hinder the search of the motif pattern and the main repeat segments.

Comparing with TRF and RepeatMasker, the main advantage of our algorithm is that, relying on the motif matrix, our algorithm will collect the information from the whole sequence which makes our algorithm to be sensitive to the case where multiple tandem repeat segments with the same motif pattern are dispersively distributed in a sequence. A detailed comparison of our algorithm with TRF and RepeatMasker, in terms of modeling and computing, is presented in Section 4.2.2 of Supplementary Materials.

DISCUSSION

Based on a matrix representation of the tandem repeat pattern, we built a probabilistic model for sequences with DSATR and introduced a *de novo* Bayesian approach to detect both the repeat pattern and the repeats locations. MCMC algorithms, including Gibbs sampler, M-H and RJMCMC, are used to estimate all parameters in the model. As to our knowledge, this article might be the first to use RJMCMC for repeat detection in biological sequences.

Although MCMC methods are appealing for exploring complex posterior distributions, it is always a concern that the MCMC chain will be trapped in some local modes. We used group moves for highly correlated parameters, such that the chains move more globally to avoid being trapped in local modes.

To tackle the unknown number of repeat segments in a full Bayesian way, we used RJMCMC to jump between models of different dimensions, which allowed the posterior probability to speak up and search for the number of repeat segments within one single MCMC chain. Two versions of our RJMCMC algorithms are introduced. Both theoretical analysis and computational experiments showed that the piloted version significantly increased the efficiency upon the vanilla version. Comparing with existing methods which can be used for DSATR identification, our RJMCMC algorithm outperforms GMS and TRF in terms of both sensitivity and specificity, and appears more sensitive than RepeatMasker for the synthetic sequences even if RepeatMasker is given with the true consensus pattern. Similar to GMS, we did not consider indels in our main algorithm, but a post-processing step for indels was applied as an auxiliary part to detect the possible repeat units with indels. The real data experiments suggest that our main algorithm is suitable for searching the main repeat region and the motif matrix of DSATR, and the post-processing step efficiently detects the nearly repeat units with indels.

Once we generalized the motif model to detect tandem repeats, many previous works on motif discovery can be adapted to address related problems in tandem repeat studies. For example, Gupta and Liu (43) and Jensen *et al.* (30) discussed how to learn the motif width from the data. One possible extension of our current DSATR detection algorithm is to relax the fixed motif width by placing a prior on w and then update w via one more RJMCMC step. Another natural extension of the current work would be the application of the algorithm on multiple input sequences which share the same tandem repeat pattern. Future studies may also consider indels in the generative model to improve the performance on this aspect.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–3, Supplementary Figures 1–2, Supplementary Material and Supplementary References [24,26,32,39–42,44–49].

FUNDING

Research grant [CUHK 400709, in part]; Research Grants Council of the Hong Kong SAR; CUHK direct grant [CUHK 2060362, in part]. Funding for open access charge: Research grant [CUHK 400709].

Conflict of interest statement. None declared.

REFERENCES

1. Tóth,G., Gáspári,Z. and Jurka,J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.*, **10**, 967–981.
2. Verstrepen,K.J., Jansen,A., Lewitter,F. and Fink,G.R. (2005) Intragenic tandem repeats generate functional variability. *Nat. Genet.*, **37**, 986–990.
3. Myers,S., Freeman,C., Auton,A., Donnelly,P. and McVean,G. (2008) A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet.*, **40**, 1124–1129.
4. Sutherland,G.R. and Richards,R.I. (1995) Simple tandem DNA repeats and human genetic disease. *Proc. Natl Acad. Sci. USA*, **92**, 3636–3641.

5. Leeflang,E.P., Zhang,L., Tavaré,S., Hubert,R., Srinidhi,J., MacDonald,M.E., Myers,R.H., de Young,M., Wexler,N.S., Gusella,J.F. *et al.* (1995) Single sperm analysis of the trinucleotide repeats in the Huntington's disease gene: quantification of the mutation frequency spectrum. *Hum. Mol. Genet.*, **4**, 1519–1526.
6. Wang,S., Wang,M., Yin,S., Fu,G., Li,C., Chen,R., Li,A., Zhou,J., Zhang,Z. and Liu,Q. (2008) A novel variable number of tandem repeats (VNTR) polymorphism containing Sp1 binding elements in the promoter of XRCC5 is a risk factor for human bladder cancer. *Mutat. Res. Fundam. Mol. Mech. Mutagen.*, **638**, 26–36.
7. Lu,Q., Wallrath,L.L., Granok,H. and Elgin,S.C. (1993) $(CT)_n(GA)_n$ Repeats and heat shock elements have distinct roles in chromation structure and transcriptional activation of the Drosophila HSP26 gene. *Mol. Cell. Biol.*, **13**, 2802–2814.
8. Du,J., Zhu,Y., Shanmugam,A. and Kenter,A.L. (1997) Analysis of immunoglobulin SGAMMA3 recombination breakpoints by PCR: implications for the mechanism of isotype switching. *Nucleic Acids Res.*, **25**, 3066–3073.
9. Weber,J. and May,P. (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain-reaction. *Am. J. Hum. Genet.*, **44**, 388–396.
10. Kimura,M., Sakamuri,R.M., Groathouse,N.A., Rivoire,B.L., Gingrich,D., Krueger-Koplin,S., Cho,S.-N., Brennan,P.J. and Vissa,V. (2009) Rapid variable-number tandem-repeat genotyping for mycobacterium leprae clinical specimens. *J. Clin. Microbiol.*, **47**, 1757–1766.
11. Moretti,T.R., Baumstark,A.L., Defenbaugh,D.A., Keys,K.M., Smerick,J.B. and Budowle,B. (2001) Validation of short tandem repeats (STRs) for forensic usage: performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples. *J. Forensic Sci*, **46**, 647–660.
12. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
13. Bao,Z. and Eddy,S.R. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.*, **12**, 1269–1276.
14. Kurtz,S., Choudhuri,J.V., Ohlebusch,E., Schleiermacher,C., Stoye,J. and Giegerich,R. (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.*, **29**, 4633–4642.
15. Kolpakov,R., Bana,G. and Kucherov,G. (2003) mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.*, **31**, 3672–3678.
16. Kurtz,S., Narechania,A., Stein,J. and Ware,D. (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genom.*, **9**, 517.
17. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*, **27**, 573–580.
18. Sussillo,D., Kundaje,A. and Anastassiou,D. (2004) Spectrogram analysis of genomes. *EURASIP J. Adv. Signal Process.*, **2004**, 29–42.
19. Tran,T.T., Emanuele,V.A. and Zhou,G.T. (2004) Techniques for detecting approximate tandem repeats in DNA. In: *Proceeding of the IEEE International Conference on Acoustic Speech Signal Process*, Vol. 5, pp. 449–452.
20. Sharma,D., Issac,B., Raghava,G.P.S. and Ramaswamy,R. (2004) Spectral repeat finder (SRF): Identification of repetitive sequences using Fourier transformation. *Bioinformatics*, **20**, 1405–1412.
21. Zhou,H.X., Du,L. and Yan,H. (2009) Detection of tandem repeats in DNA sequences based on parametric spectral estimation. *IEEE Trans. Inf. Technol. Biomed.*, **13**, 747–755.
22. Saha,S., Bridges,S., Magbanua,Z.V. and Peterson,D.G. (2008) Empirical comparison of ab initio repeat finding programs. *Nucleic Acids Res.*, **36**, 2284–2294.
23. Leclercq,S., Rivals,E. and Jarne,P. (2007) Detecting microsatellites within genomes: significant variation among algorithms. *BMC Bioinformatics*, **8**, 125.
24. Liu,J. (2001) *Monte Carlo Strategies in Scientific Computing*. Springer, New York.
25. Gelman,A., Carlin,J.B., Stern,H.S. and Rubin,D.B. (2004) *Bayesian Data Analysis*. Chapman & Hall/CRC, New York.
26. Lawrence,C., Altschul,S., Boguski,M., Liu,J., Neuwald,A. and Wootton,J. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
27. Liu,J.S., Neuwald,A.F. and Lawrence,C.E. (1995) Bayesian models for multiple local sequence alignment and gibbs sampling strategies. *J. Am. Stat. Assoc*, **90**, 1156–1170.
28. Roth,F.P., Hughes,J.D., Estep,P.W. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol*, **16**, 939–945.
29. Bailey,T. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, Vol. 1, pp. 28–36.
30. Jensen,S.T., Liu,X.S., Zhou,Q. and Liu,J.S. (2004) Computational Discovery of gene regulatory binding motifs: A bayesian perspective. *Stat. Sci.*, **19**, 188–204.
31. Li,Q., Fan,X., Liang,T. and Li,S.Y.R. (2011) A Markov chain Monte Carlo algorithm for detecting short adjacent repeats in multiple sequences. *Bioinformatics*, **27**, 1772–1779.
32. Green,P.J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.
33. Brooks,S.P., Giudici,P. and Roberts,G.O. (2003) Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *J. Roy. Stat. Soc. B*, **65**, 3–39.
34. Al-Awadhi,F., Hurn,M. and Jennison,C. (2004) Improving the acceptance rate of reversible jump MCMC proposals. *Stat. Prob. Lett.*, **69**, 189–198.
35. Stephens,M. (2000) Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Stat.*, **28**, 40–74.
36. Cappe,O., Robert,C.P. and Ryden,T. (2003) Reversible jump, birth-and-death and more general continuous time markov chain Monte Carlo samplers. *J. Roy. Stat. Soc. B*, **65**, 679–700.
37. Petris,G. and Tardella,L. (2003) A geometric approach to transdimensional Markov chain Monte Carlo. *Can. J. Stat.*, **31**, 469–482.
38. Bühlmann,P. (2000) Model selection for variable length Markov chains and tuning the context algorithm. *Ann. Inst. Stat. Math.*, **52**, 287–315.
39. Jensen,S.T. and Liu,J.S. (2004) BioOptimizer: a Bayesian scoring function approach to motif discovery. *Bioinformatics*, **20**, 1557–1564.
40. Kingman,J.F.C. (1982) The coalescent. *Stochastic Process. Appl.*, **13**, 235–248.
41. Kingman,J.F.C. (1982) On the genealogy of large populations. *J. Appl. Probab.*, **19**, 27–43.
42. Wakeley,J. (2008) *Coalescent Theory: An Introduction*. Roberts & Company Publishers, Greenwood Village, Colorado.
43. Gupta,M. and Liu,J.S. (2003) Discovery of conserved sequence patterns using a stochastic dictionary model. *J. Am. Stat. Assoc*, **98**, 55–66.
44. Liu,J.S. (1994) The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Am. Stat. Assoc*, **89**, 958–966.
45. Gelman,A. and Rubin,D.B. (1992) Inference from iterative simulation using multiple sequences. *Stat. Sci.*, **7**, 457–472.
46. Kruglyak,S., Durrett,R.T., Schug,M.D. and Aquadro,C.F. (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl Acad. Sci. USA*, **95**, 10774–10778.
47. Ellegren,H. (2004) Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.*, **5**, 435–445.
48. Ohta,T. and Kimura,M. (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res.*, **22**, 201–204.
49. Weber,J.L. and Wong,C. (1993) Mutation of human short tandem repeats. *Hum. Mol. Genet.*, **2**, 1123–1128.