A Model-Based Bayesian Estimation of the Rate of Evolution of VNTR Loci in *Mycobacterium tuberculosis*

R. Zachariah Aandahl[1,2], Josephine F. Reyes[3], Scott A. Sisson[1], Mark M. Tanaka[2]*

[1] School of Mathematics and Statistics, University of New South Wales, Sydney, New South Wales, Australia, [2] Evolution & Ecology Research Centre and School of Biotechnology & Biomolecular Sciences, University of New South Wales, Sydney, New South Wales, Australia, [3] The Kirby Institute, University of New South Wales, Sydney, New South Wales, Australia

Abstract

Variable numbers of tandem repeats (VNTR) typing is widely used for studying the bacterial cause of tuberculosis. Knowledge of the rate of mutation of VNTR loci facilitates the study of the evolution and epidemiology of *Mycobacterium tuberculosis*. Previous studies have applied population genetic models to estimate the mutation rate, leading to estimates varying widely from around 10^{-5} to 10^{-2} per locus per year. Resolving this issue using more detailed models and statistical methods would lead to improved inference in the molecular epidemiology of tuberculosis. Here, we use a model-based approach that incorporates two alternative forms of a stepwise mutation process for VNTR evolution within an epidemiological model of disease transmission. Using this model in a Bayesian framework we estimate the mutation rate of VNTR in *M. tuberculosis* from four published data sets of VNTR profiles from Albania, Iran, Morocco and Venezuela. In the first variant, the mutation rate increases linearly with respect to repeat numbers (linear model); in the second, the mutation rate is constant across repeat numbers (constant model). We find that under the constant model, the mean mutation rate per locus is $10^{-2.06}$ (95% CI: $10^{-2.61}$,$10^{-1.58}$)and under the linear model, the mean mutation rate per locus per repeat unit is $10^{-2.45}$ (95% CI: $10^{-3.07}$,$10^{-1.94}$). These new estimates represent a high rate of mutation at VNTR loci compared to previous estimates. To compare the two models we use posterior predictive checks to ascertain which of the two models is better able to reproduce the observed data. From this procedure we find that the linear model performs better than the constant model. The general framework we use allows the possibility of extending the analysis to more complex models in the future.

Citation: Aandahl RZ, Reyes JF, Sisson SA, Tanaka MM (2012) A Model-Based Bayesian Estimation of the Rate of Evolution of VNTR Loci in *Mycobacterium tuberculosis*. PLoS Comput Biol 8(6): e1002573. doi:10.1371/journal.pcbi.1002573

Editor: Sergei L. Kosakovsky Pond, University of California San Diego, United States of America

Received October 31, 2011; **Accepted** May 4, 2012; **Published** June 28, 2012

Funding: This work was supported by an Australian Postgraduate Award to RZA. MMT and SAS are supported by the Australian Research Council under the Discovery Project scheme (DP0987302 and DP1092805). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: m.tanaka@unsw.edu.au

Introduction

Mycobacterium tuberculosis, the bacterial pathogen that causes tuberculosis, latently infects one third of the world's population and is responsible for the highest mortality rate of any single bacterial pathogen [1]. Recent advances in genotyping techniques have increased our ability to discriminate among *M. tuberculosis* isolates, helping to shed light on the genetic diversity, demographics and evolution of this pathogen [2,3]. For instance, Pepperell et al. [4,5] suggested that the restricted diversity in this bacterial species is likely the result of population bottlenecks and founder effects. Genotyping or fingerprinting also refines our understanding of the epidemiological characteristics of the disease in a population, for example by revealing the extent of local transmission and factors associated with this transmission (e.g., [6]).

Frequently used methods for genetic fingerprinting of *M. tuberculosis* include restriction fragment length polymorphism typing based on mobility of the insertion sequence IS *6110* [7] and spoligotyping which exploits variation at the Direct Repeat or CRISPR locus [8]. More recently, a multilocus typing method based on variable numbers of tandem repeats (VNTR) has been developed for *M. tuberculosis* [9–11]. These loci are minisatellites, and are also known as mycobacterial interspersed repetitive units (MIRUs). We will refer to these as "VNTR loci".

VNTR-based methods are increasing in importance and efforts are being made to standardise the loci used [9]. The larger the number of loci used, the greater the discrimination among isolates resulting in a large number of smaller clusters of identical profiles in a sample. The early standard of 5 locus VNTR typing lacked the discriminatory power of IS*6110*-typing but comparative studies have shown that using at least 12 loci can have comparable or better discrimination relative to IS*6110* [12–14]. An advantage of using VNTR is that if the mutation rate is low there is the possibility of adding more loci to increase discriminatory power [10].

Inferences about transmission are sensitive to the degree of genetic clustering, which is a function of the mutation rate of the marker [15]. It is therefore important to have accurate estimates of the mutation rate of VNTR loci. Knowledge of the mutation rate of VNTR also allows calibration of the molecular clock to make inferences about the evolutionary history of *M. tuberculosis*, for instance, the time until the most recent common ancestor of a clade [3].

Author Summary

Genetically typing the bacterium responsible for tuberculosis is useful for understanding the evolutionary and epidemiological characteristics of the disease. Typing methods based on variable number tandem repeat (VNTR) loci are increasingly being used. These loci, which are composed of repeated units, mutate by increasing or decreasing in the number of these repeats. Knowledge of the mutation rate of molecular markers facilitates the epidemiological interpretation of the observed genetic variation in a sample of bacterial isolates. Few studies have examined the rate of mutation at these markers and estimates to date have varied considerably. To address this problem we develop a stochastic model of evolution of these markers and then estimate their mutation rate using approximate Bayesian computation. We examine two alternative forms of the mutation process. The observed data are from four published data sets of tuberculosis bacterial isolates sampled in Albania, Iran, Morocco and Venezuela. We find that these markers have fairly high rates of mutation compared with estimates from previous studies.

A standard model for the evolution of VNTR loci is the stepwise mutation model [16,17], which has successfully been used to describe microsatellite evolution in eukaryotes (e.g. [18]). The stepwise mutation model has also been applied to VNTR evolution in *M. tuberculosis* [19], leading to estimates of the rate of mutation. Such estimates in the literature vary widely from 10^{-5} per locus per year [19] to $10^{-3.9}$ per locus per year [3] to 10^{-3}–10^{-2} [20]. This wide variation in estimates has led to debate in the literature [21–24]. Taking a model-based approach can help to resolve this question. It allows our understanding of biological mechanisms underlying VNTR evolution to be incorporated into the analysis, while providing a natural framework for model validation and criticism. Similarly, examination of multiple data sets under the same models and methods could provide support or otherwise for resulting estimates.

In this study we estimate the mutation rate of VNTR markers by developing a stochastic stepwise mutation process of the evolution of genotypes through gains and losses of repeat numbers [16,19] embedded in a model of disease transmission [25]. We consider and evaluate two alternative formulations of the stepwise mutation model under a Bayesian statistical framework, applying our methods to four geographically distinct data sets. Our study provides a posterior estimate of the VNTR mutation rate under an explicit model of evolution placed within an epidemiological context.

Methods

Model of the dynamics of infection and mutation of VNTR loci

In the model of disease transmission we use, $S(t)$ tracks the number of individuals who are susceptible to infection and $X(t)$ tracks infectious individuals, where t is time measured in years. For simplicity, we assume a population of fixed size N. Let $\beta>0$ be the rate of transmission and $\delta>0$ be the rate of death or recovery. First consider a deterministic model where the dynamics are given by

$$\frac{dX(t)}{dt}=\beta(N-X(t))\frac{X(t)}{N}-\delta X(t). \tag{1}$$

We start the process with a single infected individual ($X(0)=1$). Define R_0 to be the basic reproductive ratio, that is, the number of cases resulting from a single infectious case in a wholly susceptible population. For this model, $R_0=\beta/\delta$. The analytical solution of Equation (1) can be written as

$$X(t)=\frac{N(R_0-1)}{[N(R_0-1)-R_0]e^{-(\delta R_0-\delta)t}+R_0}. \tag{2}$$

The steady state of the infectious population is

$$\lim_{t\to\infty}X(t)=N\left(1-\frac{1}{R_0}\right).$$

We use this deterministic model as the basis for a continuous-time stochastic model that incorporates mutation at VNTR loci. The transition rates of this model, summarised in Table 1, are as follows: the rate of new infections is $\beta(N-X(t))X(t)/N$ and the rate out of the infectious class from death or recovery is $\delta X(t)$. An infection event increases $X(t)$ by 1 while a death-or-recovery event decreases $X(t)$ by 1. Each infection is associated with a bacterial *genotype* by which we mean the set of repeat states across all loci considered in a VNTR typing technique, determined for a particular isolate. Let $X_i(t)$ be the number of individuals infected with bacterial genotype i so that

$$X(t)=\sum_{i=1}^{G(t)}X_i(t)$$

where $G(t)$ is the number of distinct genotypes in the population at time t.

We apply the stepwise mutation model to describe VNTR mutation [16,17,19] in which an event results in a unit increase or decrease in the number of repeats at a locus. We define M_i to be the mutation rate per infectious case for genotype i so that the transition rate for mutation of genotype i is $M_iX_i(t)$. A mutation event results in either a new genotype, or a pre-existing genotype in the population (i.e., homoplasy). In the event of mutation to a new genotype, the number of individuals from the mutating genotype decreases by 1 and the number of individuals in the new class becomes 1. In the case of homoplasy, the number of individuals in the mutating genotype decreases by 1 while the number of individuals in the existing class increases by 1. In either case the total number of infected cases, $X(t)$, does not change.

Table 1. Transition rates in the stochastic model.

Event	Transition	Rate
Infection	$X(t)\rightarrow X(t)+1$	$\beta(N-X(t))X(t)/N$
	$X_i(t)\rightarrow X_i(t)+1$	$\beta(N-X(t))X_i(t)/N$
Death	$X(t)\rightarrow X(t)-1$	$\delta X(t)$
	$X_i(t)\rightarrow X_i(t)-1$	$\delta X_i(t)$
Mutation	$X_i(t)\rightarrow X_i(t)-1$	$M_iX_i(t)$
	$G(t)\rightarrow G(t)+1^*$	$\sum_{i=1}^{G(t)}M_iX_i(t)$
	$X_{G(t)}(t)=1^*$	$M_iX_i(t)$

*If an existing genotype is re-created by mutation, the count of that genotype is incremented instead. Note that the increment $G(t)\rightarrow G(t)+1$ occurs before the assignment $X_{G(t)}(t)=1$.

doi:10.1371/journal.pcbi.1002573.t001

We consider two alternative ways to specify VNTR mutation. In the first model, the mutation rate at a locus is proportional to the number of repeats at that locus. In this *linear model*, the per-locus mutation rate increases linearly with the number of repeats at the locus. In the second *constant model*, the mutation rate the per-locus mutation rate is constant and thus not dependent on repeat number. Defining L to be the number of loci, $R_{i,j}$ to be the number of repeats at locus j for genotype i, and $\mu_1 > 0$ to be the rate of mutation at a locus with a single repeat, under the linear model

$$M_i = \mu_1 \sum_{j=1}^{L} R_{i,j}.$$

Under the constant model

$$M_i = \mu \sum_{j=1}^{L} 1[R_{i,j} > 0],$$

where $\mu > 0$ is the per locus mutation rate and where the indicator function $1[A] = 1$ if A is true and 0 otherwise. In both models the boundary condition $R_{i,j} = 0$ is an absorbing state in that a locus with zero repeats cannot gain or lose repeats.

The process starts at time $t = 1$ with a single infected individual and the population evolves until time $t = T_{stop}$. The initial individual has genotype given by $(\phi_1, \ldots, \phi_L)$, which we call the founding genotype. At time T_{stop} a sample of size n is taken from the population. We simulate this process using the Gillespie exact algorithm [26] so that the time between events is distributed exponentially, with parameter $\lambda(t)$, where

$$\lambda(t) = \beta(N - X(t))\frac{X(t)}{N} + \delta X(t) + \sum_{i=1}^{G(t)} M_i X_i(t).$$

Given an event, the probability of a specific outcome is proportional to the rate of that outcome, so that

$$P(\text{infection}|\text{event}) = \frac{\beta(N - X(t))X(t)/N}{\lambda(t)}$$

$$P(\text{death}|\text{event}) = \frac{\delta X(t)}{\lambda(t)}$$

$$P(\text{mutation}|\text{event}) = \frac{\sum_{i=1}^{G(t)} M_i X_i(t)}{\lambda(t)}.$$

Given a mutation event, the probability of mutation in an individual with genotype i is

$$P(\text{mutation at genotype } i|\text{mutation}) = \frac{M_i X_i(t)}{\sum_{i=1}^{G(t)} M_i X_i(t)}$$

and given a mutation event in genotype i, the probability that it occurs at locus j under the linear model is

$$P_l(\text{mutation at locus } j|\text{mutation at genotype } i) = \frac{\mu_1 R_{i,j}}{M_i},$$

and under the constant model is

$$P_c(\text{mutation at locus } j|\text{mutation at genotype } i) = \frac{\mu 1[R_{i,j} > 0]}{M_i}.$$

We assume that given a mutation event at locus j in genotype i, the probability of repeat gain is equal to the probability of repeat loss, following [3,19].

Inference procedure

We implement a standard Bayesian analysis of model parameters using approximate Bayesian computation (ABC) [27–29]. ABC methods permit approximate Bayesian inference when numerical evaluation of the posterior distribution is either computationally prohibitive or not available, and have been successfully applied to problems in molecular epidemiology [30–34].

Intuitively, given a candidate parameter vector, $\theta \in \Theta$, prior distribution $\pi(\theta)$ and model likelihood $\pi(y_0|\theta)$ with observed data y_0, ABC methods proceed by generating an artificial dataset from the model $y \sim \pi(y|\theta)$ and then reducing the dataset to a low dimensional vector of summary statistics, $s = s(y)$. If s is similar to the same vector of statistics obtained from the observed data, $s_0 = s(y_0)$, then θ could have credibly reproduced the observed data under the model. As such, the parameter vector is then retained as part of the approximate posterior, otherwise it is discarded. More precisely, the posterior obtained under ABC methods is given by

$$\pi(\theta|s_0) \approx \int K_\varepsilon(s - s_0)\pi(s|\theta)\pi(\theta)ds \qquad (3)$$

where $K_\varepsilon(u) = K(|u|/\varepsilon)/\varepsilon$ is a standard smoothing kernel with scale parameter $\varepsilon > 0$. As ε becomes small, the approximation (3) becomes increasingly accurate, although computational overheads increase. If the vector of summary statistics are informative for the model parameters, then this posterior distribution approximates the true posterior distribution so that $\pi(\theta|s_0) \approx \pi(\theta|y_0) \propto \pi(y_0|\theta)\pi(\theta)$. See e.g. [30,31,35,36] for further description of ABC methods.

The parameter vector for the constant model above is $\theta_c = \{R_0, \mu, T_{stop}, \phi_1, \ldots, \phi_L\}$ where $\phi_1, \ldots, \phi_L$ is the repeat structure of the founding genotype in the simulation. For the linear model we have $\theta_l = \{R_0, \mu_1, T_{stop}, \phi_1, \ldots, \phi_L\}$. Except where this may cause confusion, we will refer to a non-model-specific parameter vector as θ.

Conditional on the parameter vector θ, and following simulation under the model, a sample of size n individuals is drawn from the resulting population. Summary statistics, s, are then computed, determined as quantities expected to be highly informative regarding the model parameters. Using lower case letters (e.g. g, r_{ij}) to denote sample-based values of the population-level counterparts (e.g. $G, R_{i,j}$), the summary statistics include the number of distinct genotypes in the sample, g, and the set of L sample means of repeats at each locus

$$\bar{r}_{.j} = \frac{1}{n}\sum_{i=1}^{g} x_i r_{ij},$$

for $j = 1, \ldots, L$, which is expected to contain information about the initial repeat numbers $\phi_1, \ldots, \phi_L$ for some time after the founding case. Here, x_i denotes the number of individuals in the sample with genotype i, and r_{ij} denotes the within-sample number of repeats at locus j for genotype i. The final two statistics are based on the ANOVA decomposition $\mathrm{SS_{total}} = \mathrm{SS_{between}} + \mathrm{SS_{within}}$ given by

$$\sum_{j=1}^{L}\sum_{i=1}^{g} x_i(r_{ij}-\bar{r}_{..})^2 = \sum_{j=1}^{L}\sum_{i=1}^{g} x_i(\bar{r}_{.j}-\bar{r}_{..})^2 + \sum_{j=1}^{L}\sum_{i=1}^{g} x_i(r_{ij}-\bar{r}_{.j})^2,$$

where $\bar{r}_{..} = \frac{1}{nL}\sum_{j=1}^{L}\sum_{i=1}^{g} x_i r_{ij}$, from which $\mathrm{MS_{between}} = \mathrm{SS_{between}}/(\mathrm{L}-1)$ and $\mathrm{MS_{within}} = \mathrm{SS_{within}}/(\mathrm{L}(\mathrm{n}-1))$ can be computed. These two statistics are expected to be informative about the mutation rate between and within loci. The complete vector of summary statistics is then given by

$$s = \{g, \mathrm{MS_{within}}, \mathrm{MS_{between}}/\mathrm{MS_{within}}, \mathrm{r}_{.1}, \mathrm{r}_{.2}, \ldots, \mathrm{r}_{.\mathrm{L}}\}.$$

To complete the model specification, we set the parameter δ to 0.52, following [32,37]. This death/recovery rate is the sum of the death rate due to tuberculosis, the death rate due to other causes, and the recovery rate from tuberculosis. We chose an informative prior distribution for R_0 based on the study of the basic reproductive value of tuberculosis by Blower et al. [38]. We use a distribution approximating the histogram in Figure 3a in reference [38] which has a mean of 5.16 and a standard deviation of 2.82, and in particular define the prior of (R_0-1) to be a gamma distribution with a shape parameter of $(5.16-1)^2/2.82^2$ and a scale parameter of $2.82^2/(5.16-1)$. The priors for $\log_{10}(\mu)$, $\log_{10}(\mu_1)$, $\log_{10}(T_{stop})$ and $\phi_1, \ldots, \phi_L$ are uniform with wide ranges as shown in Table 2.

We examine the effectiveness of the ABC inference procedure by evaluating its ability to recover accurate estimates of the mutation rate based on data generated under the constant and linear models We simulated a population of $N=5000$ individuals with $L=24$ loci, $R_0=2$, $T_{stop}=300$, and considered a range of mutation rates under each model varying across orders of magnitude $\log_{10}(\mu) \in \{-3.5, -3, -2.5, -2\}$ and $\log_{10}(\mu_1) \in \{-4, -3.5, -3, -3\}$. The number of repeats of the founding genotype were initialised as $1_3, 2_7, 3_5, 4_4, 5_3, 6_1, 8_1$ (determined as random draws from $\mathrm{Binomial}(9, 3/11)+1$), where a_b denotes b loci with repeat number a. Based on a sample of size $n=200$ we generated data under each mutation rate value, and obtained weighted samples from the ABC posterior approximations $\pi(\theta|s)$ (c.f. 3) using a population-based ABC algorithm, following [32,39,40]. The technical algorithmic details are given in Text S1.

Table 2. Prior distributions and initial sampling distributions for each model parameter.

Parameter	Prior distribution	Initial sampling distribution
$\log_{10}(\mu_1)$	$U(-5,-1)$	$U(-4,-2)$
$\log_{10}(\mu)$	$U(-5,-1)$	$U(-4,-2)$
R_0-1	Gamma(2.18,1.91)	Gamma(2.18,1.91)
$\log_{10}(T_{stop})$	$U(-\infty,\infty)$	$U(2,4)$
ϕ_j	Uniform on $\{1,2,\ldots\}$	Binom(9,1/5)+1

Initial sampling distributions are utilised in the ABC simulations (see Text S1).
doi:10.1371/journal.pcbi.1002573.t002

The estimated posterior distributions of $\log_{10}(\mu)$ and $\log_{10}(\mu_1)$ using the simulated data are shown in Figure 1. These results indicate that mutation rates can generally be recovered accurately, with the true parameter values lying in regions of high posterior density close to the posterior mode, and with a clear location shift in the density with varying mutation rate. Higher precision can be attained by using a larger sample size, although $n=200$ already represents a sample larger than the real datasets used for this study (c.f. Table 3). In the ABC setting, posterior precision can also be improved by reducing the kernel scale parameter ε in (3) or by the inclusion of more summary statistics [30,31,35,36], although each of these can substantially increase computational overheads. Improving the precision of posterior parameter estimates for given summary statistics is currently an area of active ABC research [41].

Data

We selected recently published VNTR loci data sets from studies undertaken in four countries: Albania [42], Iran [43], Morocco [11] and Venezuela [44]. We chose data sets with a high number of isolates largely from the same clade, a high number of VNTR loci in the typing method, and relatively short periods of isolate collection. The data from Albania and Venezuela are based on 24-locus typing, and the data from Iran and Morocco are based on 15 and 12 loci respectively. A summary of these data are provided in Table 3, along with the incidence of tuberculosis for each country.

As an initial exploratory examination of these data, we computed gene diversity [45] (also known as virtual heterozygosity), for each locus in each data set. This statistic is given by $1-\sum_k (n_{jk}/n)^2$ where $n_{jk} = \sum_i 1[r_{ij}=k]x_i$ is the number of isolates with repeat size k at locus j. Figure 2 (left plots) shows the empirical cumulative distribution function of gene diversity across loci for each of the data sets. There is no obvious bimodality in these distributions. This feature is consistent with a common process generating diversity, compared to, for example, the potential bi- or multi-modality in the empirical cumulative distribution function arising from a multi-modal distribution of mutation rates. Similarly, plotting the proportion of VNTR states per locus per repeat (right plots of Figure 2) reveals that while some loci are more variable than others, there is no obvious separation between loci exhibiting high and low variation.

Results

Figure 3 shows the marginal posterior distribution of the mutation rate of VNTR loci for each of the four data sets analysed. In the case of the linear model we also show (middle panel of Figure 3) the posterior of $\bar{\mu} = \bar{r}_{..}\mu_1$, the per-locus mutation rate μ_1 at repeat size 1 scaled by the average repeat number $\bar{r}_{..}$ of each dataset to provide estimates of the mean per-locus mutation rate in a population with the same distribution of repeats as found in each sample. The posterior means of the mutation rate under the two models, along with 95% central credibility intervals are given in Table 4. The mean per-locus mutation rate at a locus with a single repeat from the four data sets under the linear model is $10^{-2.45}$, and under the constant model the mean per-locus rate is $10^{-2.06}$. Note that the prior distributions of the mutation parameters are uniform on a logarithmic (base 10) scale, and so Figure 3 displays the posterior distributions on this scale.

To evaluate the suitability of the constant and linear models to describe the observed data, we follow [36,46,47] and implement posterior predictive model checks. This approach examines the predictive distribution of specified validation statistics (based on

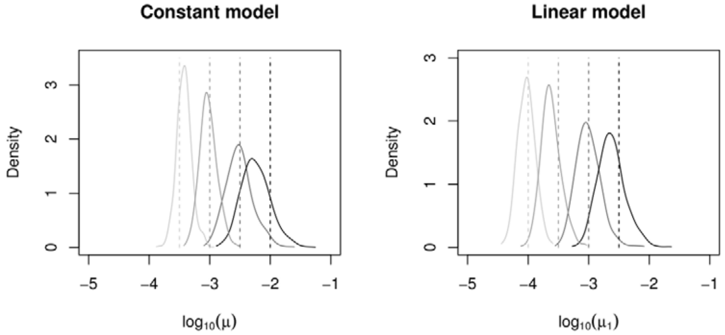

Figure 1. Marginal posterior distributions for $\log_{10}(\mu)$ and $\log_{10}(\mu_1)$ using simulated data. Plots show the marginal posterior distribution of μ (left) and μ_1 (right) using four simulated data sets generated from the constant (left) and linear (right) VNTR models. The known values of μ and μ_1 used to generate the data, $\log_{10}(\mu)\in\{-3.5,-3,-2.5,-2\}$ and $\log_{10}(\mu_1)\in\{-4,-3.5,-3,-3\}$, are indicated by vertical lines.
doi:10.1371/journal.pcbi.1002573.g001

data-generation under the fitted models) expected to be informative about various model aspects. Comparing the predictive distribution of these statistics with the same statistics derived from the observed data, enables some degree of discrimination between models. To avoid confusing model fitting with model assessment, these statistics should be different from those used in the ABC model fitting process.

Unlike the constant model, the mutation rate increases with repeat number under the linear model, and so we expect variation in repeat numbers to increase with repeat numbers. Our model assessment statistics aim to capture these differences from the data. Specifically, we focus on measures of the spread of repeats over the loci. Defining

$$\omega_1 = \max_j\{\Delta^{(j)}\}, \qquad \omega_2 = \max_j\{\Delta^{(j)}\} - \min_j\{\Delta^{(j)}\}$$

where $\Delta^{(j)} = \max_i\{r_{ij}\} - \min_i\{r_{ij}\}$, and

$$\omega_3 = \max_j\{v^{(j)}\} \qquad \text{and} \qquad \omega_4 = \max_j\{v^{(j)}\} - \min_j\{v^{(j)}\},$$

where $v^{(j)} = \sum_i x_i(r_{ij} - \bar{r}_{.j})^2/(n-1)$, and j indexes loci as before, we consider the maximum (over loci) range (ω_1), the difference between maximum and minimum range (ω_2), maximum variance (ω_3) and the difference between maximum and minimum variance (ω_4).

Under the linear model, the distributions of these statistics are expected to be shifted to higher values compared to the constant model. We also fit a simple linear regression to each data set with the standard deviation of repeat number at a locus as the response variable and the mean repeat number at a locus as the predictor variable. Based on this fit, we consider

$$\omega_5 = \text{ the regression slope } \text{ and } \omega_6 = \text{the intercept at one repeat,}$$

where ω_6 is the fitted standard deviation in repeats at a locus with a mean repeat number of one. These statistics are expected to be informative in that the slope should be positive under the linear model and near zero under the constant value, and the intercept should be low under the linear model and high under the constant model.

Figure 4 displays the predictive distributions of ω_2 versus ω_4 under both models. The observed data statistics are indicated by a cross (×). If the cross does not lie within the body of the predictive distribution, this suggests that the model and data are inconsistent with respect to aspects of the data captured by these statistics. The lower four panels present these diagnostics for artificial data generated under both models. The linear data (lower images) can be seen to be inconsistent with the constant model, but consistent with the linear model. The constant data (middle images) appear to be consistent with both models. As such, these diagnostics are able to reject the constant model when the data is generated by the linear model. In terms of the actual empirical data, the top plots in Figure 4 are based on the data from Albania. Clearly, the constant model is insufficient to describe the variation in repeat numbers inherent in the data. The linear model is better able to account for the observed pattern of repeat variation, although it is still imperfect. The posterior predictive distributions using the data sets from the other three countries were very similar to those of the Albanian data set (not shown).

The question of whether the linear model is adequate is examined further in Figure 5 which shows a posterior predictive check of ω_1 versus ω_6 under the linear model for each of the analysed data sets. In each case, the observed data lie on the periphery of the predictive densities. Although the linear model is partially able to reproduce these statistics, this analysis shows that there is room for improvement.

Table 3. Summary of data sets analysed in this study.

Country	TB incidence*	Loci	Isolates	Collection period	Source
Albania	15	24	100	2006–2007	[42]
Iran	19	15	154	2004–2005	[43]
Morocco	92	12	153	1997–1998	[11]
Venezuela	33	24	67	1997–2007	[44]

*per 100,000 per year. Data from [57].
doi:10.1371/journal.pcbi.1002573.t003

Discussion

We have analysed VNTR data from four tuberculosis studies using a model combining marker mutation and disease transmission processes, within a Bayesian framework. Our analysis shows that the VNTR mutation rate is likely to be relatively high – the posterior mean is higher than some previous estimates obtained in the literature [3,19] and closer to more recent estimates [20]. The four data sets, which are from different geographic regions, yielded very similar estimates. Such agreement of estimates is expected if there is a common mechanism of mutation across data sets.

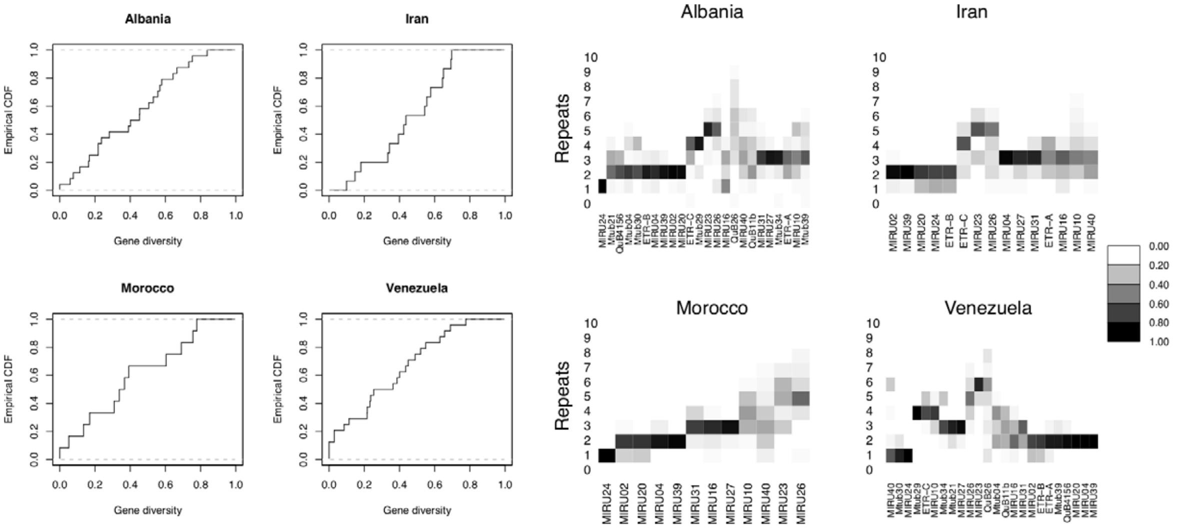

Figure 2. Genetic diversity of VNTR loci for each published dataset. Left plots: Empirical cumulative distribution function of gene diversity across loci. The gene diversity is computed at each locus as $1-\sum_k (n_{jk}/n)^2$ where $n_{jk}=\sum_i 1[r_{ij}=k]x_i$ is the number of isolates with repeat size k at locus j. Right plots: Heat-map diversity, following Aminian et al (2009), illustrating the proportion of tandem repeats for each locus (ordered according to the original study).
doi:10.1371/journal.pcbi.1002573.g002

Previous work by two of us [20] used standard equilibrium results of the infinite alleles model to describe mutation at multiple VNTR loci, and used estimates of other markers (IS6110 and spoligotyping) to calibrate the VNTR rates. That population genetic approach did not account for evolution of VNTRs as a stepwise mutation process. It therefore did not account for homoplasy, though this problem is mitigated by the inclusion of multiple VNTR loci. Further, the underlying dynamics did not include any epidemiological details. Nevertheless, it allowed us to analyse a large number of data sets in the literature to provide a ballpark estimate of VNTR mutation rates. In contrast to that and other prior work, here we used a model that explicitly and simultaneously accounts for the mutation process of the marker and the disease dynamics, and we explored two alternative models of mutation. In addition, the stepwise mutation model used here allows mutation events to re-generate existing VNTR profiles, thereby accounting for homoplasy [48].

In the debate over the magnitude of VNTR mutation rates [3,21–24] it has been noted that if loci are classified as less variable and more variable, then lower values would be estimated from the former category of loci. This raises the question of whether classification of loci into two categories of rates is supported by an underlying bimodal distribution whose modes correspond to low and high levels of polymorphism. In examining gene diversity, which is a measure of polymorphism, across loci in each data set (Figure 2) we did not observe any obvious break separating less and more variable loci. We have therefore pooled all loci and obtained an estimate of the rate of an arbitrary locus, rather than for a subset of slow or fast evolving loci. If hypermutable VNTR loci exist and are excluded from estimation procedures, using the remaining loci would clearly yield a lower mutation rate.

Our use of the linear model is a step towards resolving this issue. The linear relationship by which more units of a repeat are more

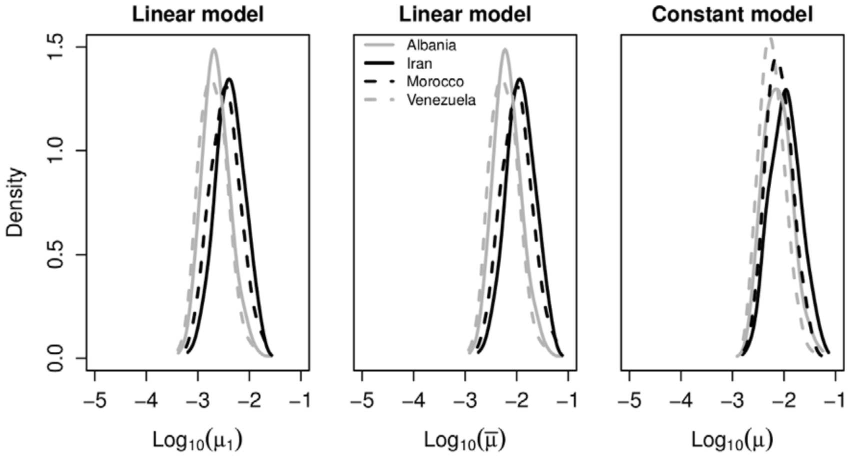

Figure 3. Marginal posterior estimates for $\log_{10}(\mu_1)$, $\log_{10}(\bar{\mu})$ and $\log_{10}(\mu)$. Here μ_1 is the per-locus mutation rate for a locus with a single repeat under the linear model; $\bar{\mu}=\bar{r}_{..}\mu_1$ is the same quantity scaled by the mean number of repeats observed in the sample; μ is the per-locus mutation rate for any repeat number under the constant model.
doi:10.1371/journal.pcbi.1002573.g003

Table 4. Bayesian posterior estimates for mutation rate.

	μ_1		μ	
Country	**mean**	**95% credible interval**	**mean**	**95% credible interval**
Albania	$10^{-2.54}$	$(10^{-3.14},10^{-2.05})$	$10^{-2.08}$	$(10^{-2.64},10^{-1.52})$
Iran	$10^{-2.37}$	$(10^{-2.94},10^{-1.82})$	$10^{-1.95}$	$(10^{-2.53},10^{-1.57})$
Morocco	$10^{-2.44}$	$(10^{-3.04},10^{-1.86})$	$10^{-2.19}$	$(10^{-2.58},10^{-1.54})$
Venezuela	$10^{-2.48}$	$(10^{-3.19},10^{-2.10})$	$10^{-2.03}$	$(10^{-2.67},10^{-1.69})$

doi:10.1371/journal.pcbi.1002573.t004

prone to mutation naturally creates variation in rates. In fact, in assessing the ability of each of our two mutation models to describe the data, we found that the linear model performs better than the constant model (Figure 4). We note that the average mutation rate $\bar{\mu}$ under the linear model was estimated to be very close to the mutation rate μ in the constant model; in this sense our analysis is robust to the exact form of the mutation model.

Despite the linear model outperforming the constant model, a posterior predictive goodness-of-fit analysis revealed some evidence that the linear model did not fit the data perfectly (Figure 5).

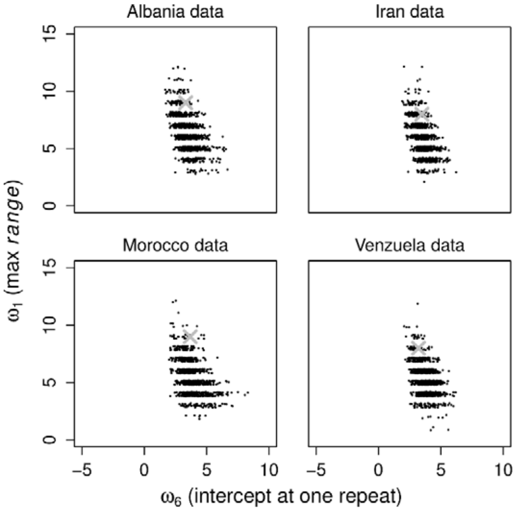

Figure 5. Further posterior predictive model checks. Scatterplots of the posterior predictive distributions of ω_1 (the maximum range of repeat numbers over loci) versus ω_6 (the intercept at one repeat) under the linear model, for each observed dataset. The × indicates the statistics derived from the observed dataset.
doi:10.1371/journal.pcbi.1002573.g005

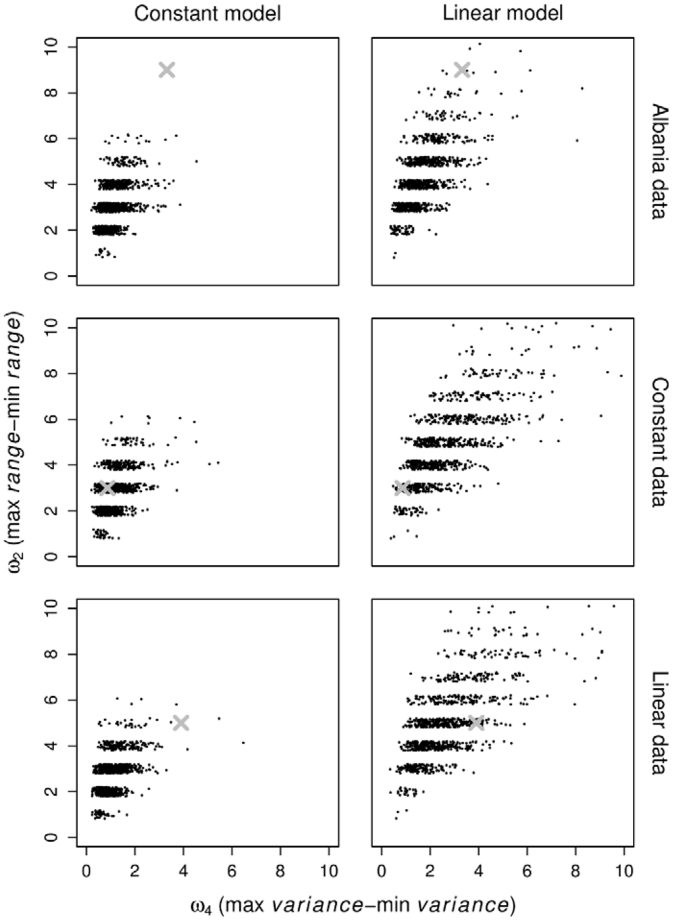

Figure 4. Posterior predictive model checks. Scatterplots of the posterior predictive distributions of ω_2 (the difference between maximum and minimum range of repeat numbers over loci), versus ω_4 (the same quantity substituting variance for range). Columns represent constant (left) and linear (right) models. Rows represent the Albanian dataset (top), artificially generated data from the constant model (middle) and artificially generated data from the linear model (bottom). The × indicates the statistics derived from the observed dataset.
doi:10.1371/journal.pcbi.1002573.g004

While previous studies of eukaryote minisatellites agree with a linear relationship between repeat number and mutation rate [49], some studies of eukaryote microsatellites indicate a more complex relationship between repeat number and mutation rate [50–53]. We investigated a third model in which the mutation rate increases exponentially with repeat number, but the results are very similar to those of the linear model (Figure S3 in Text S1). Future work might adopt a per locus mutation rate that grows non-linearly with repeat number. A drawback of this possibility would be the added complexity and dimensionality of the model with the need to estimate further parameters in a framework that is already computationally intensive. An alternative approach might be to construct a hierarchical Bayesian model of mutation rates in which each locus is associated with its own rate according to some distribution, akin to the analysis of Bazin et al. [54].

We have used a simple model to avoid overfitting the data. However, it is possible to extend the model in future studies to incorporate further complexity and realism. One such detail is the reactivation of latent infection, which could be described by a susceptible-exposed-infected (SEI) model in which a proportion of cases progress directly to disease [38]. We performed preliminary simulations from a stochastic version of such a model (details in Text S1). We consider the number of distinct genotypes since this is one of the statistics we use in the inference and it is known to be informative for mutation rate in similar models [55,56]. Figure S2 in Text S1 shows how the number of distinct genotypes in a sample varies with the mutation rate under both models. The latent reactivation model was able to generate statistics close to the observed statistic. The points in the region of the observed statistic are near the posterior density generated under the original model. While this is suggestive that a latency model would produce similar estimates, a full Bayesian analysis would be required to address this issue. The lack of latency is a limitation of our study which should be addressed in future research.

Migration is another factor which a more realistic multi-deme population model might incorporate. The interplay between migration and mutation may affect the resulting estimates of the mutation rate. For example, migration from regions with genetically very different clades of *M. tuberculosis* occurs at a high rate would lead to over-estimation of the mutation rate. Our approach based on the approximate Bayesian computation framework makes future directions such as this and those relating to the mutation process feasible.

Supporting Information

Text S1 Additional technical details of the algorithm used in the Bayesian analysis, the stochastic model of latent tuberculosis reactivation, and the mutation model of VNTR with an exponential increase in rate with respect to repeat number.
(PDF)

Acknowledgments

We thank anonymous reviewers for their suggestions.

Author Contributions

Conceived and designed the experiments: RZA JFR SAS MMT. Performed the experiments: RZA. Analyzed the data: RZA SAS MMT. Contributed reagents/materials/analysis tools: RZA JFR. Wrote the paper: RZA JFR SAS MMT.

References

1. World Health Organization (2010) Global tuberculosis control 2010. World Health Organization.
2. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, et al. (2008) High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. PLoS Biol 6: e311.
3. Wirth T, Hildebrand F, Allix-Béguec C, Wölbeling F, Kubica T, et al. (2008) Origin, spread and demography of the *Mycobacterium tuberculosis* complex. PLoS Pathog 4: e1000160.
4. Pepperell C, Hoeppner V, Lipatov M, Wobeser W, Schoolnik GK, et al. (2010) Bacterial genetic signatures of human social phenomena among M. tuberculosis from an Aboriginal Canadian population. Mol Biol Evol 27: 427–440.
5. Pepperell CS, Granka JM, Alexander DC, Behr MA, Chui L, et al. (2011) Dispersal of *Mycobacterium tuberculosis* via the Canadian fur trade. Proc Natl Acad Sci U S A 108: 6526–6531.
6. Sails AD, Barrett A, Sarginson S, Magee JG, Maynard P, et al. (2011) Molecular epidemiology of *Mycobacterium tuberculosis* in East Lancashire 2001–2009. Thorax 66: 709–713.
7. Thierry D, Brisson-Noel A, Vincent-Levy-Frebault V, Nguyen S, Guesdon JL, et al. (1990) Characterization of a *Mycobacterium tuberculosis* insertion sequence, IS6110, and its application in diagnosis. J Clin Microbiol 28: 2668–2673.
8. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, et al. (1997) Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol 35: 907–914.
9. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rusch-Gerdes S, et al. (2006) Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. J Clin Microbiol 44: 4498–4510.
10. Oelemann MC, Diel R, Vatin V, Haas W, Rüsch-Gerdes S, et al. (2007) Assessment of an optimized mycobacterial interspersed repetitive- unit-variable-number tandem-repeat typing system combined with spoligotyping for population-based molecular epidemiology studies of tuberculosis. J Clin Microbiol 45: 691–697.
11. Tazi L, El Baghdadi J, Lesjean S, Locht C, Supply P, et al. (2004) Genetic diversity and population structure of *Mycobacterium tuberculosis* in Casablanca, a Moroccan city with high incidence of tuberculosis. J Clin Microbiol 42: 461–466.
12. Sola C, Filliol I, Legrand E, Lesjean S, Locht C, et al. (2003) Genotyping of the *Mycobacterium tuberculosis* complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics. Infect Genet Evol 3: 125–133.
13. Valcheva V, Mokrousov I, Narvskaya O, Rastogi N, Markova N (2008) Utility of new 24-locus variable-number tandem-repeat typing for discriminating *Mycobacterium tuberculosis* clinical isolates collected in Bulgaria. J Clin Microbiol 46: 3005–3011.
14. Smittipat N, Billamas P, Palittapongarnpim M, Thong-On A, Temu MM, et al. (2005) Polymorphism of variable-number tandem repeats at multiple loci in *Mycobacterium tuberculosis*. J Clin Microbiol 43: 5034–5043.
15. Tanaka MM, Francis AR (2005) Methods of quantifying and visualising outbreaks of tuberculosis using genotypic information. Infect Genet Evol 5: 35–43.
16. Ohta T, 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.
17. Shriver MD, Jin L, Chakraborty R, Boerwinkle E (1993) VNTR allele frequency distributions under the stepwise mutation model: A computer simulation approach. Genetics 134: 983–993.
18. Cornuet JM, Beaumont MA, Estoup A, Solignac M (2006) Inference on microsatellite mutation processes in the invasive mite, varroa destructor, using reversible jump Markov chain Monte Carlo. Theor Popul Biol 69: 129–144.
19. Grant A, Arnold C, Thorne N, Gharbia S, Underwood A (2008) Mathematical modelling of *Mycobacterium tuberculosis* VNTR loci estimates a very slow mutation rate for the repeats. J Mol Evol 66: 565–574.
20. Reyes JF, Tanaka MM (2010) Mutation rates of spoligotypes and variable number tandem repeat loci in *Mycobacterium tuberculosis*. Infect Genet Evol 10: 1046–1051.
21. Supply P, Niemann S, Wirth T (2011) On the mutation rates of spoligotypes and variable numbers of tandem repeat loci of *Mycobacterium tuberculosis*: Continued when tuning matters. Infect Genet Evol 11: 1191–1191.
22. Supply P, Niemann S, Wirth T (2011) On the mutation rates of spoligotypes and variable numbers of tandem repeat loci of *Mycobacterium tuberculosis*. Infect Genet Evol 11: 251–252.
23. Tanaka MM, Reyes JF (2011) Mutation rate of VNTR loci in *Mycobacterium tuberculosis*: Response to Supply et al. Infect Genet Evol 11: 1189–1190.
24. Tanaka MM, Reyes JF (2011) VNTR mutation in *Mycobacterium tuberculosis*: Lower rates for less variable loci. Infect Genet Evol 11: 1192–1192.
25. Jacquez JA, Simon CP (1993) The stochastic SI model with recruitment and deaths I. Comparison with the closed SIS model. Math Biosci 117: 77–125.
26. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81: 2340–2361.

27. Tavare S, Balding DJ, Grifiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. Genetics 145: 505–518.
28. Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. Genetics 162: 2025–2035.
29. Marjoram P, Molitor J, Plagnol V, Tavaré S (2003) Markov chain Monte Carlo without likelihoods. Proc Natl Acad Sci U S A 100: 15324–15328.
30. Beaumont MA (2010) Approximate Bayesian computation in evolution and ecology. Ann Rev Ecol Sys 41: 379–406.
31. Bertorelle G, Benazzo A, Mona S (2010) ABC as a exible framework to estimate demography over space and time: some cons, many pros. Mol Ecol 19: 2609–2625.
32. Luciani F, Sisson SA, Jiang H, Francis AR, Tanaka MM (2009) The epidemiological fitness cost of drug resistance in *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A 106: 14711–14715.
33. Tanaka MM, Francis AR, Luciani F, Sisson SA (2006) Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. Genetics 173: 1511–1520.
34. Drovandi CC, Pettitt AN (2011) Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. Biometrics 67: 225–233.
35. Sisson SA, Fan Y (2011) Likelihood-free Markov chain Monte Carlo. In: Brooks SP, Gelman A, Jones G, Meng XL, editors. Handbook of Markov chain Monte Carlo. Chapman and Hall/CRC Press. pp. 319–341.
36. Csillery K, Blum MGB, Gaggiotti OE, Francois O (2010) Approximate Bayesian computation (ABC) in practice. Trends Ecol Evol 25: 410–418.
37. Cohen T, Murray M (2004) Modeling epidemics of multidrug-resistant M. *tubercu-losis* of heterogeneous fitness. Nat Med 10: 1117–1121.
38. Blower S, Mclean A, Porco T, Small P, Hopewell P, et al. (1995) The intrinsic transmission dynamics of tuberculosis epidemics. Nat Med 1: 815–821.
39. Sisson SA, Fan Y, Tanaka MM (2007) Sequential Monte Carlo without likelihoods. Proc Natl Acad Sci U S A 104: 1760–1765. Errata (2009), 106, 16889.
40. Peters GW, Fan Y, Sisson SA (2012) On sequential Monte Carlo, partial rejection control and approximate Bayesian computation. Stat Comput : in press.
41. Fearnhead P, Prangle D (2012) Constructing summary statistics for approximate Bayesian computation: Semi-automatic ABC. J Roy Stat Soc B 74: 419–474.
42. Tafaj S, Zhang J, Hauck Y, Pourcel C, Hafizi H, et al. (2009) First insight into genetic diversity of the *Mycobacterium tuberculosis* complex in Albania obtained by multilocus variable-number tandem-repeat analysis and spoligotyping reveals the presence of Beijing multidrug-resistant isolates. J Clin Microbiol 47: 1581–1584.
43. Asgharzadeh M, Khakpour M, Salehi TZ, Kafil HS (2007) Use of mycobacterial interspersed repetitive unit-variable-number tandem repeat typing to study *Mycobacterium tuberculosis* isolates from East Azarbaijan province of Iran. Pak J Biol Sci 10: 3769–3777.
44. Abadía E, Sequera M, Ortega D, Méndez M, Escalona A, et al. (2009) *Mycobacterium tuberculosis* ecology in Venezuela: epidemiologic correlates of common spoligotypes and a large clonal cluster defined by MIRU-VNTR-24. BMC Infect Dis 9: 122–134.
45. Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci U S A 70: 3321–3323.
46. Gelman A, Meng XL, Stern HS (1996) Posterior predictive assessment of model fitness via realized discrepancies. Stat Sinica 6: 733–807.
47. Thornton K, Andolfatto P (2006) Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics 172: 1607–1619.
48. Reyes JF, Chan CHS, Tanaka MM (2012) Impact of homoplasy on variable numbers of tandem repeats and spoligotypes in *Mycobacterium tuberculosis*. Infect Genet Evol 12: 811–818.
49. Buard J, Bourdet A, Yardley J, Dubrova Y, Jeffreys A (1998) Inuences of array size and homogeneity on minisatellite mutation. EMBO J 17: 3495–3502.
50. Dieringer D, Schlötterer C (2003) Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Res 13: 2242–2251.
51. Lai Y, Sun F (2003) The relationship between microsatellite slippage mutation rate and the number of repeat units. Mol Biol Evol 20: 2123–2131.
52. Kelkar Y, Tyekucheva S, Chiaromonte F, Makova K (2008) The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res 18: 30–38.
53. Seyfert A, Cristescu M, Frisse L, Schaack S, Thomas W, et al. (2008) The rate and spectrum of microsatellite mutation in *Caenorhabditis elegans* and *Daphnia pulex*. Genetics 178: 2113–2121.
54. Bazin E, Dawson KJ, Beaumont MA (2010) Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. Genetics 185: 587–602.
55. Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3: 87–112.
56. Blum MGB, Nunes MA, Prangle D, Sisson SA (2012). A comparative review of dimension reduction methods in approximate Bayesian computation. http://arxiv.org/abs/1202.3819.
57. World Health Organization (2009). Tuberculosis Country Profiles: Epidemiology and Strategy. http://www.who.int/tb/country/data/profiles/en/index.html.