AUSTRIAN JOURNAL OF STATISTICS
Volume 36 (2007), Number 1, 53–64

Hierarchical Models for Mitochondrial DNA Sequence Data

Paola Berchialla
University of Torino, Italy

Abstract: We introduce a Bayesian hierarchical model for mitochondrial DNA sequence data, which is fitted via acceptance-rejection algorithms. The model incorporates parametric models of population history explicitly as well as a mutational process allowing for a simultaneous parameter estimation whose importance has become increasingly clear in many recent studies. The model is applied to a sample of DNA sequences from the Italian population.

Abstract: Wir stellen ein Bayes'sches hierarchisches Modell für mitochondrialen DNA Sequenzdaten vor, das mittels Acceptance-Rejection Algorithmen angepasst wird. Das Modell enthält explizit parametrische Modelle für die Entwicklung der Population wie auch einen Veränderungsprozess und erlauben eine simultane Parameterschätzung. Die Wichtigkeit dieser Vorgehensweise ist durch viele aktuelle Studien ganz deutlich nachgewiesen worden. Das Model wird auf eine Stichprobe von DNA Sequenzen der Italienischen Population angewandt.

Keywords: Bayesian Modelling, Coalescent Process, Acceptance-Rejection Algorithm.

1 Introduction

Inference about population histories and evolutionary processes are not only of intrinsic interest but are also crucial to the interpretation of genetic data in a wide range of applications which vary from molecular biology and genetic medicine to forensic sciences (see Bataille et al., 1999; Foreman et al., 1997; Jorde et al., 2001).

The control region (sometimes referred as D-loop) of the mitochondrial genome is widely used in studies of human population to address question concerning genetic variation within species. This is due to the maternal inheritance of mitochondrial DNA, the absence of recombination and the high mutation rate (Cann et al., 1987).

Human control region sequences evolve according to a complex pattern that makes analysis difficult. In fact underlying a sample of DNA sequences data is a structure shaped by dependencies that reflect the ancestral relationships between the sequences and are affected by historical patterns of migration, mating behavior and population growth, as well as mutation and selection (Cavalli-Sforza et al., 1994). In the absence of recombination, these relationships can be represented by a genealogical tree for which each leaf corresponds to a sequence at the present time while the root of the tree represents the most recent common ancestor of all the sequences in the sample (Wilson et al. (2003)).

Although the underlying relationships are crucial in modelling the dependence structure of a sample of DNA sequences, they are in fact ignored by traditional methods, most of which are based on the distribution of pairwise differences, obtained by comparing

pairs of sequences and recording the number of pairs with $0, 1, \ldots$ differences, or summary statistics like the number of segregating sites, i.e. the number of single positions or loci in a sequence which experienced a mutation (see Bonneuill, 1998; Tavaré et al., 1997).

In recent years important advances have been made in developing tree reconstruction methods and computational techniques such as Markov chain Monte Carlo and importance sampling (Stephens, 2001). Tree reconstruction methods can give insights into the mode of evolution of the genomic region studied. In particular, coalescent theory provides a framework with which to incorporate parametric models of population history explicitly. Coalescent models, which describe the evolution of a sample of DNA sequences in terms of stochastic processes, allow application of statistical techniques for parameter estimation and model testing (Stephens and Donnelly, 2000).

The trade-off is that the computational complexity of the analysis can increase substantially and implementing these models and algorithms is challenging. Furthermore they assume a constant rate at which mutations occur while an important feature of mitochondrial sequence evolution is the variation of rates among sites. To gauge the contribution of hot spot, i.e. positions at which substitutions accumulate predominantly, to the high rate estimate, it is necessary to infer the rate for each polymorphic locus in the sequence (see Wakeley, 1994; Swofford et al., 1995).

The aim of this paper is to introduce a Bayesian hierarchical model to estimate demographic and mutational parameters of a sample of mitochondrial DNA sequences. In order to take into account the underlying ancestral relationships between sequences we use coalescent models as prior distributions.

In the next section we give a brief outline of the standard coalescent process and the coalescent with population growth and then we introduce the hierarchical model. In Section 3 we discuss the choice of the prior distributions along with the results obtained by the application of the model to a sample from Italian population.

2 Theory and Methods

2.1 Standard Coalescent Process

The coalescent is a stochastic model for the genealogical tree representing the ancestral relationships between a sample of DNA sequences. It approximates the distribution of genealogical trees under an important class of neutral population genetics models, including the celebrated Wright-Fisher model of a random-mating population of constant size N (see Hudson, 1990; Kingman, 1982). To recover this approximation, 1 unit of 'coalescent' must be interpreted as N generations, where N is the effective population size, say the number of adults in a population contributing offspring to the next generation (Hartl and Clark, 1997).

Coalescences occur only between pairs of individuals. This process may be thought of as generating a binary tree, with leaves representing the sample sequences and vertices where ancestral lines coalesce.

Coalescent time runs backward, with time $t_0 \equiv 0$ denoting the present and time t_k, $k \in \{1, 2, \ldots, n-1\}$, denoting the time of the k-th most recent coalescent event which

take place between n individuals. In particular, t_{n-1} denotes the time of the most recent common ancestor (TMRCA) of the sample.

The between-coalescent intervals $W_k = t_{n-k+1} - t_{n-k}$ during which the sample has k distinct ancestors have independent exponential distributions

$$\Pr(t_{n-k+1} > t | t_{n-k} = t') = \exp\left(\gamma_k(t' - t)\right) , \qquad \gamma_k = \binom{k}{2} \tag{1}$$

for $t > t'$. An important quantity associated to a coalescent tree is the height, defined as

$$T = \sum_{k=2}^{n} W_k \, .$$

By definition the expectations of W_k are equal to $2/k(k-1)$ and so the expectation of T is given by

$$\mathrm{E}(T) = 2\left(1 - \frac{1}{n}\right)$$

which approaches 2 units of coalescent time, equivalent to $2N$ generations as the sample size gets large.

2.2 Coalescent with Population Growth

The effect of variable population size is to change the joint distribution of the times t_k. Suppose that the population size at the time of sampling is N. The population size at time $N \int_0^t \lambda(s)ds$ generations ago will be written as $N\lambda(t)$. The standard coalescent model is a special case with $\lambda(s) \equiv 1$ of the model in which equation (1) is replaced by

$$\Pr(t_k > t | t_{k-1} = t') = \exp(\gamma_k(\Lambda(t') - \Lambda(t))) \, , \tag{2}$$

where $\Lambda(t) \equiv \int_0^t ds/\lambda(s)$ (Hudson, 1990).

Let us consider pure exponential growth at rate R, i.e.

$$\lambda(t) = \exp(-Rt) \, . \tag{3}$$

Since $t_{n-k} = W_{k+1} + \cdots + W_n$, using (3) in equation (2) gives

$$\Pr(W_k > s | W_{k+1} + \cdots + W_n = t') = \exp\left\{-\gamma_k \frac{\exp(Rt')}{R}\left[\exp(Rs) - 1\right]\right\} , \tag{4}$$

where $s = t - t'$.

A more realistic scenario is a two parameter model for which

$$\lambda(t) = \begin{cases} \exp(R(t_g - t)) & \text{if } \ 0 < t < t_g \\ 1 & \text{if } \ t > t_g \end{cases}$$

corresponding to a population of constant size N until Nt_g generations ago, after which it grows at rate R per generation to reach its current size N_c, where

$$N_c \approx N \exp(Rt_g) \, .$$

Under this model the coalescent time distribution (2) becomes

$$\Pr(W_k > s | W_{k+1} + \cdots + W_n = t') =$$

$$\begin{cases} \exp\{\frac{\gamma_k}{R}[\exp(Rt') - \exp(R(t'+s))]\exp(-Rt_g)\} & \text{if } t' < t'+s < t_g \\ \exp\{\gamma_k(t_g - (t'+s) + \frac{1}{R}(\exp[R(t'-t_g)] - 1])\} & \text{if } t' < t_g < t'+s \\ \exp\{-\gamma_k s\} & \text{if } t_g < t' < t'+s\,. \end{cases}$$

2.3 A Hierarchical Model

Mitochondrial DNA sequences data are a string of letters each of ones denotes the allele, i.e. the possible state of the DNA sequence at a locus. Letter A stay for adenine, C for cytosine, G for guanine, and T for thymine. However, it is a common situation that a locus exhibits just two variants across individuals, for example $T \Leftrightarrow C$ or $A \Leftrightarrow G$. Then for each locus we arbitrarily choose and fix one of the two variants as usual in single-nucleotide polymorphism allele frequencies models (Nicholson et al., 2002).

We consider the setting in which we have a collection of L loci from a modern population. Suppose n_j is the number of individuals typed at the j-th locus in the population. At the lowest levels of the hierarchical model, the number of copies x_j of the chosen variant at locus j has a binomial distribution

$$x_j \sim \text{Binomial}(n_j, \alpha_j)\,, \qquad j = 1, 2, \ldots, L$$

where $\alpha_j \in [0, 1]$ is the probability of the chosen variant at locus j.

We model the dependence structure between the modern population and the ancestral population to that sampled through a mutational and a demographic process. First introduce a collection of unobserved quantities, one for each locus: π_j, $j = 1, 2, \ldots, L$ which plays the role of the allele frequencies in the ancestral population. Then

$$\beta_j = (1 - \exp(-\mu_j t))\pi_j$$

is the probability of a mutation after tN generations (see Weir, 1990), whereas

$$\alpha_j = (1 - \exp(-\mu_j tN))\pi_j + (1 - \pi_j)\exp(-\mu_j tN) \tag{5}$$

is the probability of changes between the two variants in the j-th locus. Finally to complete the hierarchy we put independent priors on π, μ, N, t. The choice of the prior distribution is discussed in the results section.

3 Simulation Methods

In this section we describe a simulation approach which is based on the acceptance-rejection method (see Ripley, 1987) to generate observations from the posterior distribution of demographic and mutational parameters.

We assume that, before observing the data, N and μ are mutually independent random quantities and independent of coalescent times t and the ancestral population allele frequencies π. The prior distributions of N and μ should be chosen so as to summarize

the information available, for example from relevant genetic and anthropological studies. Typically, such information will not uniquely specify the distribution, therefore it is prudent to consider several different plausible choice and investigate the sensitivity of conclusions.

Algorithm 1

1. simulate N and μ from their specified prior distribution;
2. simulate π_j and T from a symmetric Beta(p) and an Exponential(2) distribution, respectively;
3. calculate α_j according to the equation (5);
4. keep (π_j, T, N, μ) with probability u defined by

$$u = \frac{\text{Binomial}(y, \alpha_j)}{\text{Binomial}(y, y/n)}\,. \tag{6}$$

When considering the coalescence time T the algorithm can be modified as follows

Algorithm 2

1. simulate N and μ from their specified prior distribution;
2. simulate π_j from a symmetric Beta(p) distribution and the W_k as independent Exponential$(k(k-1)/2)$ variables, $k = 1, 2, \ldots, n$;
3. evaluate T by $T = \sum_{k=2}^{n} W_k$;
4. calculate α_j according to equation (5);
5. keep (π_j, T, N, μ) with probability u defined by (6).

When allowing for population growth, the above algorithm can be readily modified in the following one.

Algorithm 3

1. simulate R from its prior distribution;
2. for $k = n, n-1, \ldots, 2$ simulate $t_n - k + 1$ by

$$t_{n-k+1} = \frac{1}{R}\log\left[\exp(Rt_{n-k}) - \frac{R}{\gamma_k}\log U_k\right], \qquad t_0 = 0\,;$$

   and evaluate $T = \sum_{k=2}^{n} W_k = t_{n-1}$;
3. calculate $N(t) = N_0 \exp(-RT)$, where N_0 is the current population size;
4. simulate μ and π;
5. keep (π_j, T, N, μ) with probability u defined by (6).

Finally the preceding algorithm can be modified as follows to allow for the demographic scenario provided by the two-parameter exponential growth.

Algorithm 4

1. simulate R and then simulate time t_g which corresponds to the time at which the population starts its exponential growth and N by

$$\log\left(\frac{N_c}{N}\right) = Rt_gN\,; \tag{7}$$

2. for $k = n, n-1, \ldots, 2$ simulate $t_n - k + 1$ by

$$t_{n-k+1} = \begin{cases} \frac{1}{R}\log\left[\exp(Rt_{n-k}) - \frac{R}{\gamma_k}\exp(Rt_g)\log U\right] & \text{if } t_{n-k} < t_{n-k+1} < t_g \\ t_g + \frac{1}{R}\left[\exp(R(t_{n-k} - t_g)) - 1\right] - \frac{1}{\gamma_k}\log U & \text{if } t_{n-k} < t_g < t_{n-k+1} \\ \sim \text{Exponential}(1/\gamma_k) & \text{if } t_g < t_{n-k} < t_{n-k+1} \end{cases}$$

   and evaluate $T = \sum_{k=2}^{n} W_k = t_{n-1}$;

3. simulate μ and π from their prior distributions;

4. keep $(\pi_j, T, N, N_c, R, \mu)$ with probability u defined by (6).

All the simulation routines were written using the `R` programming language (R Development Core Team, 2006).

4 Results

A data set consisting of mitochondrial DNA sequences from a sample of 49 Italian individuals was collected to illustrate the hierarchical model described above. Data are available on the FBI Mitochondrial DNA Population Database (Monson et al., 2002). Each sequence is composed of the first 360 base pair segment of the control region, corresponding to positions 16024-16383 in the human reference sequence of Anderson et al. (1981).

In the mtDNA sequences, 28 polymorphic loci which showed variation across individuals were identified. The settings of loci was considered evolving independently. Such condition typically arises unless loci are very close on the same molecule.

For setting up prior distribution in the standard coalescent model with a constant population size, we refer to Tavaré et al. (1997) They estimate the effective population size N to be of order 5000 individuals. To allow uncertainty a gamma prior distribution with mean 5000 and shape 5 was considered (Table 1). Such distribution concentrates largely on values between 0 and 6000 and it is approximately centered around the value 4900 which is also supported by Hammer (1995) and Fullerton et al. (1994). Alternatively a log-normal distribution with mean and standard deviation equal to 9 and 1 respectively was considered (see Wilson et al., 2003). Since the resulting posterior distributions were similar, results were summarized only for the posterior distribution obtained by assuming the gamma prior in Table 2.

In the coalescent model with exponential population growth, uncertainty of growth parameter R was modelled by means of a gamma distribution with mean 0.005, shape 2

Table 1: Demographic parameters from prior distribution for algorithms 1 and 2

Parameters	1st Qu.	Median	Mean	3rd Qu.
		Prior		
(a) Standard coalescent model-algorithm 1				
effective population size	3408	4677	5017	6320
TMRCA ($\times$ years $\times$ generation)	6.05^4	1.51^5	2.48^5	3.27^5
(b) Standard coalescent model-algorithm 2				
effective population size	3354	4644	4980	6246
TMRCA ($\times$ years $\times$ generation)	1.23^5	1.99^5	2.47^5	3.15^5

Table 2: Demographic parameters from posterior distribution for algorithms 1 and 2

Parameters	1st Qu.	Median	Mean	3rd Qu.
		Posterior		
(a) Standard coalescent model-algorithm 1				
Effective population size	3405	4733	5095	6370
TMRCA ($\times$ years $\times$ generation)	6.57^4	1.64^5	2.74^5	3.51^5
(b) Standard coalescent model-algorithm 2				
Effective population size	3257	4562	4866	6101
TMRCA ($\times$ years $\times$ generation)	1.3^7	1.99^7	2.23^7	2.97^7

Table 3: Demographic parameters from prior distribution for algorithm 3

Parameters	1st Qu.	Median	Mean	3rd Qu.
		Prior		
Ancestral population size	1676	2702	2983	3889
Modern population size	2.52^4	6.97^4	153^3	166.4^3
Growth rate R (% per generation)	0.24	0.41	0.49	0.67
TMRCA ($\times$ years $\times$ generation)	202.8^3	560.9^3	123.1^4	133.8^4

Table 4: Demographic parameters from posterior distribution for algorithm 3

Parameters	1st Qu.	Median	Mean	3rd Qu.
		Posterior		
Ancestral population size	1823	2750	3088	3996
Modern population size	18970	54670	143000	146700
Growth rate R (% per generation)	0.24	0.43	0.51	0.7
TMRCA ($\times$ years $\times$ generation)	174.7^3	380.7^3	607.8^3	736.4^3

and rate 400. According to Wilson et al. (2003) it was reasonable to assume an effective population size N distributed as a gamma with mean 3000 and shape and rate parameters equal to 3 and 10^{-3} respectively, while for the ancestral population size N_c, parameter $\log(N_c/N)$ was modelled as a gamma distribution with shape and rate equal to 5 and

Table 5: Demographic parameters from prior distribution for algorithm 4

Parameters	1st Qu.	Median	Mean	3rd Qu.
		Prior		
Ancestral population size	1676	2702	2983	3889
Modern population size	1.66^4	2.67^4	2.95^4	3.85
Growth rate R (% per generation)	0.24	0.41	0.49	0.67
Time since start growth	8519	1.39^4	2.25^4	2.35^4
TMRCA ($\times$ years $\times$ generation)	1.04^6	1.75^6	2.34^6	2.93^6

Table 6: Demographic parameters from posterior distribution for algorithm 4

Parameters	1st Qu.	Median	Mean	3rd Qu.
		Posterior		
Ancestral population size	429	576	747	763
Modern population size	2484	4212	6348	6764
Growth rate R (% per generation)	0.223	0.398	0.478	0.623
Time since start growth	6.71^3	1.19^4	2.61^4	2.86^4
TMRCA ($\times$ years $\times$ generation)	1.40^5	2.47^5	5.11^5	5.33^5

1, respectively. From the relationship 7 along with the assumption that $\log(N_c/N)$, N, and R were mutually independent (see Wilson et al., 2003) the distribution of time t_g at which population started its growth was derived. According to considerations as in Tavaré et al. (1997), uncertainty of the mutation rate μ_j was given in the form of a $\text{gamma} \sim (2, 4.94 \times 10^{-8})$ for all j (see Tables 3 and 5). A value of 25 years was assumed for the generation time (Hein, 2004).

Finally a prior distribution on allele frequencies in the ancestral population was taken to be a symmetric beta and alternatively a uniform distribution (Nicholson et al., 2002). Sensitivity analysis showed that conclusions did not depend on the choice of the prior distribution, thus results in Table 7 were summarized for the choice of a symmetric beta with parameter $p = 0.1$, only.

In Table 2, the estimated TMRCA using algorithm 2 was two order of magnitude higher than TMRCA estimated via algorithm 1. This is mainly due to the effect of mutations which stretch out the intervals between coalescence times.

From Table 7, positions 16146-16173-16190-16194-16257 showed a mean mutational rate which was lesser than other positions both for algorithm 3 and algorithm 4 revealing a mutation rate heterogeneity of the mitochondrial DNA control region.

To test the ability of the inference procedure introduced and recover accurate estimates of parameters a simulation study was undertook. We simulated 50 data sets from algorithm 2. The parameters underlying each simulation were obtained via the independent prior distributions discussed above.

In Table 8 let χ be the indicator function which assume value equal 1 if the $100p\%$ posterior interval includes the correct value, 0 otherwise. The observed average of χ over the data sets formed a Binomial(50, p) proportion. Since there were not appreciable

Table 7: Mutational parameters from posterior distribution for algorithms 3 and 4

Position	Algorithm 3				Algorithm 4			
	1st Qu.	Median	Mean	3rd Qu.	1st Qu.	Median	Mean	3rd Qu.
16052	5.747^{-8}	9.627^{-8}	1.129^{-7}	1.499^{-7}	5.838^{-8}	9.788^{-8}	1.141^{-7}	1.545^{-7}
16070	5.318^{-8}	9.152^{-8}	1.091^{-7}	1.463^{-7}	5.313^{-8}	9.655^{-8}	1.116^{-7}	1.508^{-7}
16146	4.907^{-8}	8.391^{-8}	1.015^{-7}	1.369^{-7}	4.788^{-8}	8.572^{-8}	1.005^{-7}	1.367^{-7}
16173	4.743^{-8}	8.021^{-8}	9.865^{-8}	1.330^{-7}	4.626^{-8}	8.079^{-8}	9.829^{-8}	1.329^{-7}
16190	5.028^{-8}	8.727^{-8}	1.018^{-7}	1.370^{-7}	4.911^{-8}	8.309^{-8}	1.009^{-7}	1.349^{-7}
16194	4.913^{-8}	8.416^{-8}	1.010^{-7}	1.383^{-7}	4.788^{-8}	8.235^{-8}	9.758^{-8}	1.314^{-7}
16224	5.066^{-8}	9.258^{-8}	1.100^{-7}	1.523^{-7}	5.340^{-8}	9.478^{-8}	1.092^{-7}	1.492^{-7}
16257	5.246^{-8}	8.873^{-8}	1.059^{-7}	1.394^{-7}	4.778^{-8}	8.396^{-8}	1.007^{-7}	1.383^{-7}
16279	5.332^{-8}	9.204^{-8}	1.104^{-7}	1.470^{-7}	5.517^{-8}	9.139^{-8}	1.105^{-7}	1.479^{-7}
16312	5.565^{-8}	9.717^{-8}	1.131^{-7}	1.508^{-7}	5.330^{-8}	9.452^{-8}	1.110^{-7}	1.486^{-7}
16363	5.548^{-8}	9.462^{-8}	1.120^{-7}	1.493^{-7}	5.366^{-8}	9.286^{-8}	1.101^{-7}	1.467^{-7}

differences results were summarized for TMRCA, effective population size N and just for one mutational and one ancestral allele frequencies parameter. The number of data sets for which $100p\%$ posterior interval included the true parameter value was slightly lesser than the mean, however the differences did not exceed three standard deviation. Ancestral population allele frequencies π_j were the parameters with the greatest differences among achieved and nominal coverage.

Table 8: Simulation study consisting of 50 data sets from the algorithm 2

$p\%$	Parameters			
	N	$\mu(16070)$	T	$\pi(16070)$
	χ-mean (SD%)			
10	3 (2.12)	4 (2.12)	3 (2.12)	2 (2.12)
30	12 (3.20)	17 (3.20)	16 (3.20)	10 (3.20)
50	27 (3.54)	23 (3.54)	19 (3.54)	19 (3.54)
70	35 (3.20)	29 (3.20)	31 (3.20)	29 (3.20)
90	40 (2.12)	46 (2.12)	38 (2.12)	39 (2.12)

5 Discussion

We fitted a hierarchical model for mitochondrial DNA sequences data by using a fully Bayesian approach implemented via acceptance-rejection algorithms. A feature of these algorithms is that they are usually efficient in term of time consuming. Otherwise, when the acceptance probability is very small the algorithm can be computational expensive and alternatively the model may be fitted via Markov chain Monte Carlo methods. How-

ever, the method described allows much more flexibility to explore the effects of different modelling assumptions. It is straightforward to adapt the algorithms given to include other demographic scenarios such as coalescent with population splitting (see Wilson et al., 2003). Furthermore the model itself allows for different sample size across loci.

Several authors (see Tavaré et al., 1997; Wilson and Balding, 1998) have implemented Bayesian methods in order to make inferences on the mutation rate μ and the effective population size N separately while, in the absence of other information, the statistical features of the sequence data are affected by the value of the compound mutation parameter $2N\mu$ (Yang, 1996). On the other hand demographic parameter estimates along with site-specific rate estimates could be used to refine models of molecular structure allowing for a better understanding of the forces and mechanisms that affect sequence evolution, Wakeley (1994). The proposed hierarchical model is thus motivated by the importance of simultaneous parameter estimation, especially whether mutational rate heterogeneity exists among sites. In fact, it allows to get an estimation of the substitution rate at the nucleotide level pointing out those sites which are more likely to experiencing a mutation.

Results obtained from the implementation of the model are consistent with other studies. In fact, the estimated value of the effective population size in the standard coalescent model (Table 2) is about 5000 which is consistent with the results of Jorde et al. (2001). On the other hand TMRCA estimated using algorithms 3 and 4, see Tables 4 and 6, is consistent with the coalescence time estimated in the European population using the last intron of the ZFX gene (Jaruzelska et al., 1999).

Recent studies (see Denver et al., 2000; Heyer et al., 2001) reveal a mitochondrial substitution rate that is two orders of magnitude higher than previous indirect estimates. This is mainly due to multiple mutations affecting coding function as well as mutational hotspots. Denver et al. (2000) gives an overall mutation rate equal to 1.6×10^{-7} per site per generation $(\pm 3.1 \times 10{-}8)$ which support results summarized in Table 7, except for position 16173 which is a slow site accordingly to Heyer et al. (2001).

An interesting generalization of this approach relies on the fact that it naturally handles missing data and could be extended to incorporate correlations between loci.

With regard to the simulation study, from Table 8 should be observed that ancestral population allele frequencies π_j are those parameters with the greatest differences among achieved and nominal coverage. This is probably due to the fact we simulated them using a symmetric beta distribution which assign the same probability to rare and common variant. Anyway this seem unlikely to have a large effect on inferences since for most parameters of interest there is a good agreement between the predicted and the actual coverage.

A limitation of this work concerns the sampling strategy. In our analysis we supposed that the sample was indeed "random". In practice this is difficult to arrange and the sensitivity of estimates and inferences to non-random sampling should be quantified.

References

Anderson, S., Bankier, A. T., Barell, B. G., De Bruijn, M. H. L., and Coulson, A. R. (1981). Sequence and organization of the human mitochondrial genome. *Nature*,

290, 457-465.

Bataille, M., Crainic, K., Leterreux, M., Durigon, M., and de Mazancourt, P. (1999). Multiplex amplification of mitochondrial DNA for human and species identification in forensic evaluation. *Forensic Science International*, *99*, 165-170.

Bonneuill, N. (1998). Population paths implied by the mean number of pairwise nucleotide differences among mitochondrial DNA sequences. *Annals of Human Genetics*, *62*, 61-73.

Cann, R. L., Stoneking, M., and Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature*, *325*, 31-36.

Cavalli-Sforza, L. L., Menozzi, A., and Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton: Princeton University Press.

Denver, D. R., Morris, K., Lynch, M., Vassilieva, L. L., and Thomas, W. K. (2000). High direct estimate of the mutation rate in the mitochondrial genome of Caenorhabditis elegans. *Science*, *289*, 2342-2344.

Foreman, L. A., Smith, A. F. M., and Evett, I. W. (1997). Bayesian analysis of DNA profiling data in forensic identification applications. *Journal of the Royal Statistical Society, A*, *160*, 429-469.

Fullerton, S. M., Harding, R. M., Boyce, A. J., and Clegg, J. B. (1994). Molecular and population genetics analysis of allelic sequence diversity at the human beta-clobin locus. *Proceedings of the National Academy of Science of the USA*, *91*, 1805-1809.

Hammer, M. F. (1995). A recent common ancestry for human Y chromosomes. *Nature*, *378*, 376-378.

Hartl, D. L., and Clark, A. G. (1997). *Principles of Population Genetics* (3rd ed.). Sunderland: Sinauer.

Hein, J. (2004). Human evolution: pedigrees for all humanity. *Nature*, *431*, 562-6.

Heyer, E., Zietkiewicz, E., Rochowski, A., Yotova, V., Puymirat, J., and Labuda, D. (2001). Phylogenetic and familial estimates of mitochondrial substitution rates: Study of control region mutations in deep-rooting pedigrees. *American Journal of Human Genetics*, *69*, 1113-1126.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. In D. J. Futuyama and J. D. Antonovics (Eds.), *Oxford Surveys in Evolutionary Biology* (Vol. 7, p. 1-44). New York: Oxford University Press.

Jaruzelska, J., Zietkiewicz, E., Batzer, M., Cole, D. E. C., Moisan, J. P., Scozzari, R., et al. (1999). Spatial and temporal distribution of the neutral polymorphisms in the last ZFX intron: Analysis of the Haplotype structure and genealogy. *Genetics*, *152*, 1091-1101.

Jorde, L. B., Watkins, W. S., and Bamshad, M. J. (2001). Population genomics: a bridge from evolutionary history to genetic medicine. *Human Molecular Genetics*, *10*, 2199-2207.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications*, *13*, 225-248.

Monson, K. L., Miller, K. W. P., Wilson, M. R., DiZinno, J. A., and Budowle, B. (2002). The mtDNA population database: an integrated software and database resource for forensic comparison. *Forensic Science International*, *4*. (http://www.fbi.gov/hq/lab/fsc/backissu/april2002/miller1.htm)

Nicholson, G., Smith, A. V., Jonsson, F., Gustaffson, O., Stefansson, K., and Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society, B*, *64*, 695-715.

R Development Core Team. (2006). *R: A language and environment for statistical computing*. Vienna, Austria. (ISBN 3-900051-07-0)

Ripley, B. D. (1987). *Stochastic Simulation*. New York: Wiley.

Stephens, M. (2001). Inference under the coalescent. In D. J. Balding, C. Cannings, and M. Bishop (Eds.), *Handbook of Statistical Genetics.* Chichester: Wiley.

Stephens, M., and Donnelly, P. (2000). Inference in molecular population genects (with discussion). *Journal of the Royal Statistical Society, B*, *62*, 605-635.

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1995). Accomodating rate heterogeneity among sites. In D. M. Hillis, C. Moritz, and B. K. Mable (Eds.), *Molecular Systematics.* Sunderland: Sinauer.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, *145*, 505-518.

Wakeley, J. (1994). Substitution rate variation among sites and the estimation of transition bias. *Molecular Biology and Evolution*, *11*, 436-442.

Weir, B. S. (1990). *Genetic Data Analysis*. Sunderland: Sinauer.

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

Wilson, I., Weale, M. E., and Balding, D. J. (2003). Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *Journal of the Royal Statistical Society, A*, *166*, 1-33.

Yang, Z. (1996). Statistical properties of a DNA sample under the finite-site model. *Genetics*, *144*, 1941-1950.

Author's address:

Paola Berchialla
Department of Public Health and Microbiology
University of Torino
10126 Torino, Italy

E-mail: `berchialla@econ.unito.it`