Dipartimento di Statistica
"Giuseppe Parenti"

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

WORKING PAPER 2002/18

Coalescence times and forensic identification problems

Paola Berchialla,
Federico M. Stefanini

Università degli Studi
di Firenze

Applied Statistics

Coalescence times and forensic identification problems

P. Berchialla*, F. M. Stefanini
Department of Statistics "G. Parenti", University of Florence

1. INTRODUCTION

The maternally inherited mitochondrial DNA (mtDNA) has been widely used to infer on human population history. There is a growing interest in the application of mtDNA in forensic identification problems, partially due to the possibility of typing sequences from very small or degraded biological samples (BATAILLE et al., 1999; LEVIN et al., 1999).

Looking at the genealogical relationship of a sample of genes is a powerful way to study human genetic diversity and coalescent theory provides a reference scheme to model genealogies. Coalescence has been also used to evaluate match probabilities in forensic identification problems (WILSON et al., 2000).

We are concerned with the following question: how much information about identification is carried by coalescence times? A typical data set consists of sequences of mtDNA from several individuals and the number of different sites between all pairs of individuals often suffices to summarize most of the information contained in the data (SLATKIN and HUDSON, 1991). The distribution of pairwise differences, or mismatch distribution, counting the number of site differences between each pair of sequences in a sample may be investigated at this purpose.

A mtDNA is often assumed to be a collection of completely linked sites (recombination is negligible), thus infinitely many sites coalescent model (ICM) is suited for the analysis of such DNA data (GRIFFITHS, 1989). Following the ICM model, new segregating sites arise from a mutation process in which new mutations do not occur at previously mutated sites.

We build on standard coalescent agreements to develop a model suited to forensic identification problems and to illustrate the procedure we have considered a data set of 27 sequences sampled from the italian population (Mitochondrial DNA Concordance database).

In Section 2 we recall the standard coalescent model which is the theoretical framework of our work. Section 3 deals with the mismatch distribution and the related distribution of pairwise coalescence times. The latter distribution may contain a considerable amount of information suited to forensic identification problems, as the results in Section 4 seem to suggest.

2. STANDARD COALESCENT MODEL

Suppose we have a sample of n individuals from a population of constant size. The basic strategy of the coalescent approach is to trace genealogies backward in time. All the n individuals may have distinct parents in the previous generation or some of them may share a common parent. Let's consider the number of distinct ancestor τ generations ago, $\tau = 0, 1, 2, \ldots$ As τ increases the number of distinct ancestors of the sample decreases until all n individuals at generation 0 share only one common ancestor (MRCA).

If two individuals pick the same parent in the previous generation we say that they coalesce. Coalescent events are assumed to occur only between pairs of individuals. This assumption is motivated by the expression that defines the probability that k individuals will have no common ancestor in the previous generation

$$\prod_{i=1}^{k-1} \frac{N-i}{N} = 1 - \frac{\binom{k}{2}}{N} + O\left(\frac{1}{N^2}\right). \tag{1}$$

in which the $O\left(\frac{1}{N^2}\right)$ term includes coalescence events of three or more individuals eventually involving different parents.

* *Address for correspondence*: P. Berchialla, Dipartimento di Statistica "G. Parenti", Università di Firenze, Viale Morgagni 59, I–50134, Florence, Italy. E–mail:berchial@ds.unifi.it

From (1) by independence of coalescence events in different generations, the probability that k individuals don't share a common ancestor in the previous τ generations is

$$\left[\prod_{i=1}^{k-1} 1 - \frac{i}{N}\right]^{\tau} \tag{2}$$

If we express time in units of N generations, that is by substituting $\tau = Nt$, then

$$\left[\prod_{i=1}^{k-1} 1 - \frac{i}{N}\right]^{Nt} = \left[\left(1 - \frac{\binom{k}{2}}{N} + O\left(\frac{1}{N^2}\right)\right)^{N}\right] \longrightarrow \exp\left(-\binom{k}{2}\right) \tag{3}$$

as $N \to \infty$, i.e. a well known diffusion approximation is obtained. A larger class of models is approximated if the coalescence time is measured in units of $N\sigma^{-2}$, where σ^2 is the variance of the number of offsprings for one individual. By assuming $\sigma^2 = 1$ we obtain the Wright–Fisher model (GRIFFITHS and TAVARÈ, 1994).

Let T_k be the amount of time during which there are k lineages. From (3) and the Wright–Fisher model, the distribution of T_k is exponential with scale parameter equal to $k(k-1)/2$.

The coalescence of n individuals may be represented by a binary tree in which leaves are the observed sequences and inner vertices are generated by coalescence events.

Since only pairs of individuals coalesce, in a sample of size n there are $n-1$ coalescent events. In the coalescent tree the $n-1$ branch lengths are given by the waiting times between successive coalescence events so by the $n-1$ coalescence times $\{T_n, T_{n-1}, \ldots, T_2\}$. Two important quantities associated with a coalescent tree are: the height T which is the time to the most recent common ancestor and the length L which is the total of all branch lengths. These quantities are defined by:

$$T = \sum_{k=2}^{n} T_k, \qquad L = \sum_{k=2}^{n} k T_k \tag{4}$$

The topology of the tree, that is the map defining pairs of coalescing lineages, completes the description of the genealogy. It's usual to represent the topology as a family of equivalence relations

$$\mathcal{R} = \{\mathcal{R}_N(\tau) : \tau = 0, 1, 2, \ldots\} \tag{5}$$

where for any two individuals i and j, we say that $i\mathcal{R}_N(\tau)j$ if and only if i and j have a common ancestor at τ generation; under the assumption of neutrality all lineages are equally likely to coalesce. It's trivial to verify $\mathcal{R}_N(\tau)$ is an equivalence relation. Clearly $\mathcal{R}_N(0) = \{\{1\}, \ldots, \{n\}\}$ and if $i\mathcal{R}_N(\tau)j$ then $i\mathcal{R}_N(\tau + h)j$, $h = 1, 2, \ldots$

Under mild conditions, the genealogical process $\{\mathcal{R}_N([Nt]) : t \geq 0\}$, where time is re–scaled in continuous time units of N and $[Nt]$ is the integer part of Nt, in case of a large population (i.e. large N) is well approximated by the n–coalescent $\{\mathcal{R}(t) : t \geq 0\}$ which is the continuous–time Markov chain with infinitesimal generator

$$Q_{\xi,\eta} = \begin{cases} -\binom{|\xi|}{2} & \text{if } \eta = \xi \\ 1 & \text{if } \xi \prec \eta \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where ξ and η are equivalence relations on $\{1, 2, \ldots, n\}$, $|\xi|$ is the number of equivalence classes of ξ and $\xi \prec \eta$ denote that η can be obtained from ξ by merging two equivalence classes of ξ into a single class of η.

Summarizing the genealogy of a sample can be described in terms of its topology and branch lengths. The topology $\mathcal{R}$ can be represented using equivalence classes for ancestor; the branch lengths $\{T_2, \ldots, T_n\}$ are given by the waiting times between successive coalescence events. So we can indicate a coalescent tree as $\{T_2, \ldots, T_n, \mathcal{R}\}$

According to the Wright–Fisher model, since neutral variants don't affect reproductive success, it is possible to separate the neutral mutation process from the genealogical process. Let $\mathcal{G} = \{T_2, \ldots, T_n, \mathcal{R}\}$ a genealogy, then mutation can be added afterwards according to a Poisson process with constant rate $\theta/2$ where $\theta = 2N\mu$ in which μ is the mutation rate per gene per generation. If the data are DNA sequences, then μ is equal to the sequence length $\times$ the mutation rate per site per generation.

3. METHOD

As noted by SLATKIN and HUDSON (1991) and ROGERS and HARPENDING (1992), the geometric distribution

$$P(k) = \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^k, \qquad \theta = 2N\mu \tag{7}$$

doesn't fit the observed pairwise differences. Equation (7) is useful for independent pairs of sequences (WATTERSON, 1975) but pairs of genes in a single sample or from the same population are correlated since they have a common history (BALL, NEIGEL and AVISE, 1990). Their history is represented by the underlying genealogy which is just one. SLATKIN and HUDSON (1991) and ROGERS and HARPENDING (1992) conclude that if the observed distribution of pairwise differences is close to a Poisson, then it is consistent with the hypothesis that the population from which those genes were sampled has been growing exponentially in size. Their results also suggest that a history of exponential growth in population size tends to force coalescent events to occur in a relatively restricted range of times. In that case, correlations between coalescence times created by the underlying genealogy are relatively unimportant.

The relationship between the distribution of pairwise differences and the distribution of coalescence times is obtained by noting that, for a given coalescence time $T_2 = t$, the number of mutations that occurred follows a Poisson distribution with mean $2\mu t$. So in the case of a constant size population, by standard coalescent theory we know

$$(S_2|T_2 = t) \sim \text{Poisson}(\theta t) \tag{8}$$

where S_2 is the number of segregating sites in a sample of two individuals and T_2 is the time to the most recent common ancestor of the same sample.

Under the infinitely many sites model, all of the information in the two sequences is captured in the number of segregating sites S_2 and, under the coalescent model, the coalescence time T_2 has an exponential distribution with mean 1. After observing $S_2 = k$, the distribution of T_2 is a Gamma with shape $1 + k$ and scale $1/(1 + \theta)$ (TAJIMA, 1983) whose density is

$$f(t|S_2 = k) = \frac{(1+\theta)^{1+k}}{k!} t^k \mathrm{e}^{-(1+\theta)t} \tag{9}$$

Then, if S_2 is the pairwise differences random variable and it is distributed like a Poisson with parameter λ, the distribution of the corresponding coalescence times is

$$P(T_2 \le t) = \sum_k P(T_2 \le t|S_2 = k)P(S_2 = k) = \mathrm{e}^{-(1+\theta)t-\lambda} \sum_k \frac{(1+\theta)^{1+k}}{k!} \frac{(\lambda t)^k}{k!} \tag{10}$$

Under the hypothesis of fluctuations in the population, the size $N(t)$, function of the time, must be considered. By standard diffusion approximation of the coalescent theory, the probability that the first coalescence event about two lineages in the interval of time $(t, t + dt)$ is (SLATKIN and HUDSON, 1991)

$$P(T_2 \le t)dt = \frac{1}{N(t)} \exp\left[-\int_0^t \frac{1}{N(s)} ds\right] dt \tag{11}$$

If we assume that the effective population size has been growing exponentially at a constant rate r, then

$$N(t) = N\mathrm{e}^{-rt} \tag{12}$$

where N is the population size before the expansion and equation (4) reduce to

$$P(T_2 \le t)dt = \frac{\mathrm{e}^{rt}}{N} \exp\left(-\frac{\mathrm{e}^{rt} - 1}{Nr}\right) dt \tag{13}$$

which is a special case of the Gumbel distribution.

From (9) we have:

$$f(t|S_2 = k) \propto f_{T_2}(t)P(S_2 = k|T_2 = t) = \frac{\mathrm{e}^{-(\theta-r)t}}{N} \exp\left(-\frac{\mathrm{e}^{rt}-1}{Nr}\right)\frac{(\theta t)^k}{k!} \tag{13}$$

If $Nr \gg 1$ then coalescence events tend to occur in a restricted range of times as we can see in figure 1.

As a consequence of this fact, correlations between coalescence times are not significative. Now if we assume $S_2 \sim \text{Poisson}(\lambda)$

$$P(T_2 \leq t) = \sum_k P(T_2 \leq t|S_2 = k)P(S_2 = k) \tag{15}$$

from which

$$f(t) = \sum_k f(t|K)P(S_2 = k) = \frac{\mathrm{e}^{-(\theta+\lambda-r)t}}{N} \exp\left(-\frac{\mathrm{e}^{rt}-1}{Nr}\right)\sum_k \frac{(\theta t\lambda)^k}{k! \cdot k!} \tag{16}$$

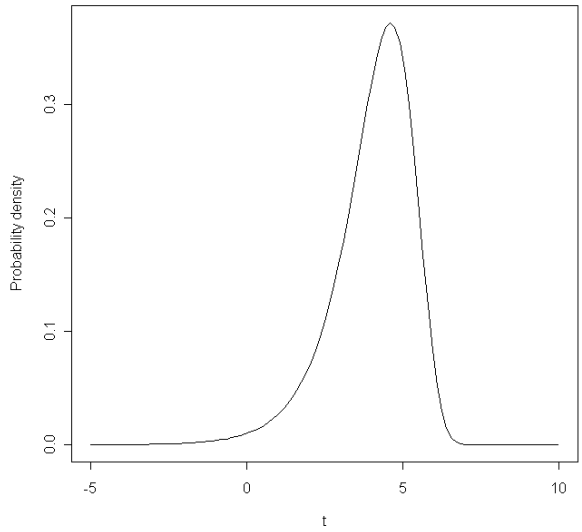

FIGURA 1. Probability of density of coalescence time, measured in units of 1/r, of two genes sampled from an exponentially growing population with Nr=100

We guess that the distribution of coalescence times T_2 obtained above contains information on the population which is referred to. In particular, it can be use as a reference distribution for samples of size two.

For example, let's A be an individual pertaining to a population $\mathcal{P}$ whose reference distribution of coalescence times is known and suppose we've found the mtDNA trace of another individual B. If A and B pertain to the same population $\mathcal{P}$ we guess that: *(i)* the mode of the distribution of coalescence time between A and B after observing the number of segregating sites doesn't differ very much from the mode of the reference distribution of population $\mathcal{P}$; *(ii)* the distribution of coalescence times between A and B given the observed number of segregating sites is quite similar to the reference distribution of population $\mathcal{P}$; whereas it is reasonable to be expected that guesses *(i)* and *(ii)* are not completely true when A and B don't pertain both to the population $\mathcal{P}$.

4. RESULTS

In figure 2 is showed the frequency distribution of pairwise differences in 27 mtDNA sequences from italian population sequenced in the HVRII 77–150 region. We have used a value of 5000 for the effective

population size N and a value of 1/300 for the mutation rate per gene per generation (STONEKING et al., 1992), corresponding to $\theta = 10.09$.

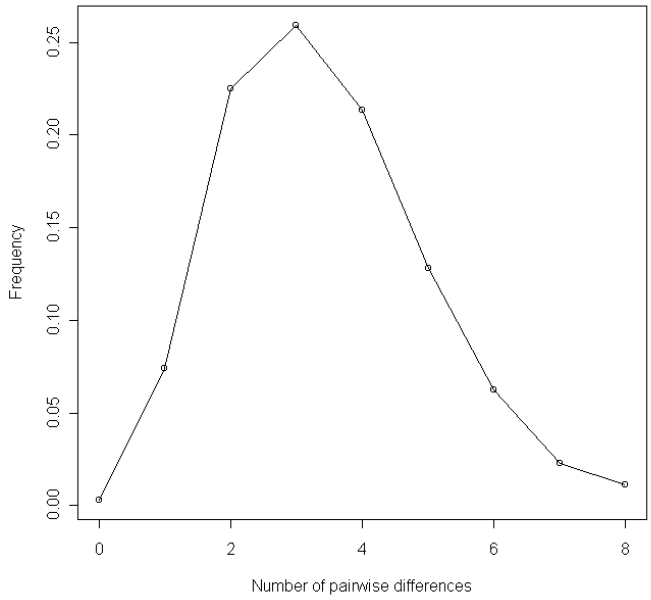

FIGURA 2. Frequency of pairwise differences in 27 mtDNA sequences

The distribution is unimodal and it strongly resembles to a Poisson distribution as regards its shape.

In figure 3 we have fitted observations from a Poisson distribution: the – – distribution is a Poisson with the same mean of the observed pairwise differences whereas the –·– distribution is a Poisson whose parameter λ is the median of the posterior $p(\lambda|k) = p(k|\lambda)p(\lambda)$ with a prior $p(\lambda)$ choose as Gamma(α, β) prior whose mean and variance are the same of the observed pairwise differences.

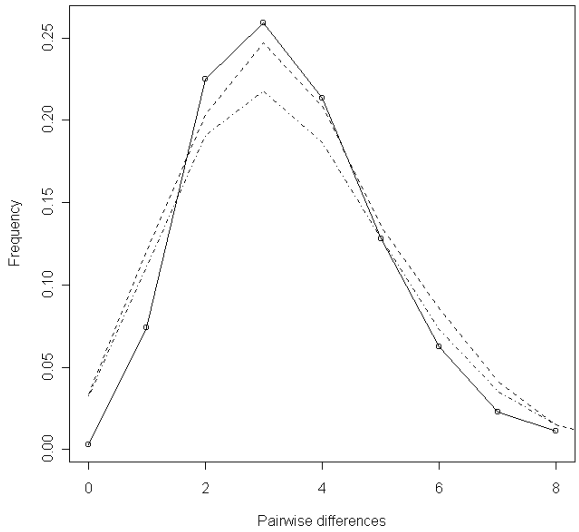

FIGURA 3. Comparison to the observed pairwise differences with a Poisson distribution.

Since the observed mismatch distribution is close to a Poisson and quite similar as regards its shape, the hypothesis of an approximatively exponential increase in population size is supported. In figure 4 we have plotted the coalescence times T_2 in the case of a constant population size, obtained as described in the previous section, and corresponding to the two Poisson distributions of figure 3. They are very similar with a coalescence time mode equals to 0.27 for the first distribution and 0.29 for the second.

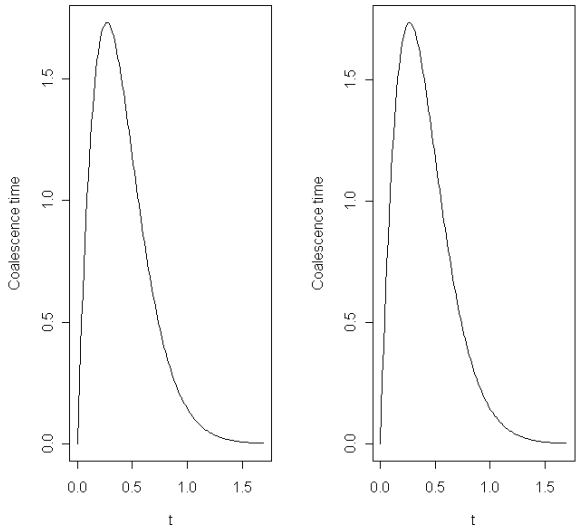

FIGURA 4. Distribution of coalescence times. For the first distribution we have marginalized on a Poisson with the same mean of the observed pairwise differences; for the second we have marginalized on a Poisson with parameter equal to the median of the posterior.

By making the hypothesis of exponentially growing population, the distribution of coalescence becomes (figure 5):

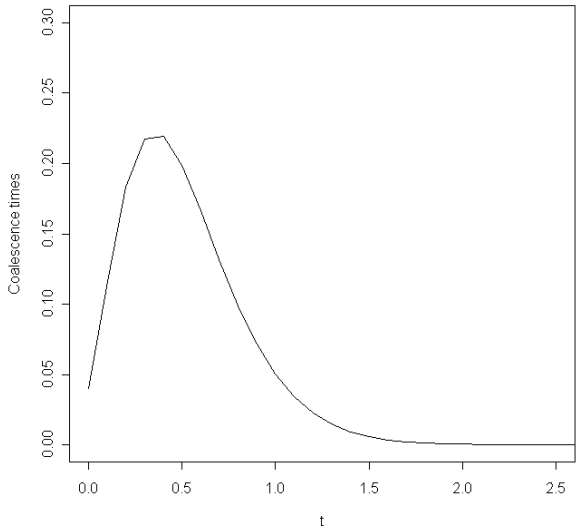

FIGURA 5. Distribution of coalescence times when population increases exponentially in size.

We are interesting in using information about coalescence times in forensic problems. Suppose we have found at the scene of crime genetic material and there is reason to believe the sample is from the perpetrator of the crime. Which is the information carried by coalescence times? If the two samples are from the same population or from the same racial group, then they should coalesce before time that occurs when they pertain to different populations. For example, the distribution of coalescence times between two italian individuals is similar to the reference distribution for the italian population we have found. In particular the mode of the coalescence times distributions is very close to the mode of reference distribution as showed in figure 6 for constant size population and in figure 7 for exponentially growing size population.

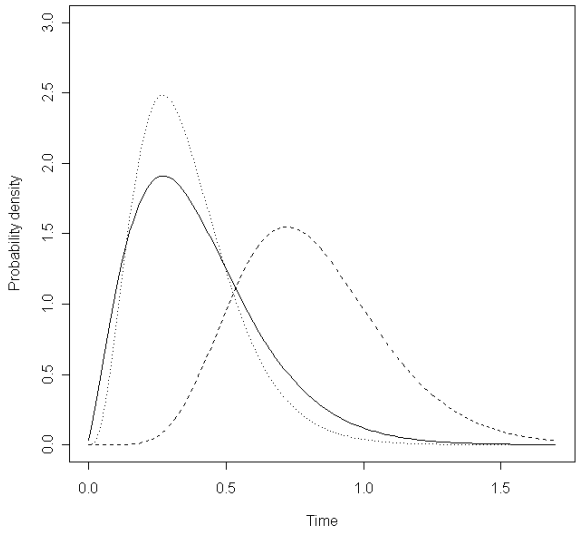

FIGURA 6. Constant size population. — reference probability density of italian population;· · · coalescence times probability density between two Italians; - - - coalescence times probability density between an italian individual and an african individual

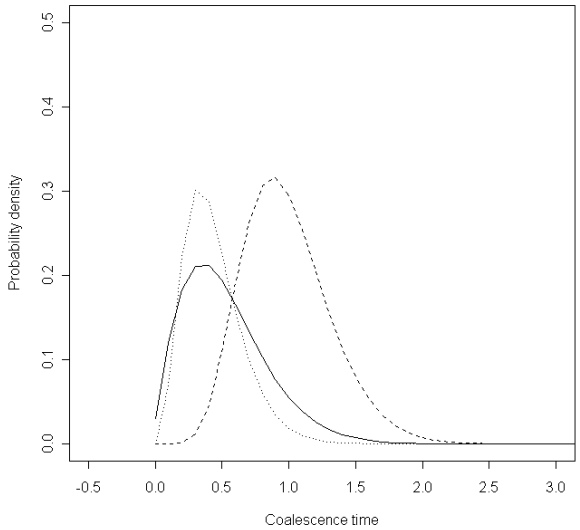

FIGURA 7. Exponentially growing population. — reference probability density of italian population; · · · coalescence times probability density between two Italians; - - - coalescence times probability density between an italian individual and an african individual

4. CONCLUSIONS

Frequency distribution of pairwise differences in sequences of genes, or equivalently their pairwise coalescence times, provides an indication of the history of individuals from which they are sampled. In particular, distributions of pairwise differences which are similar in shape to a Poisson indicate that demographic events in the past have forced coalescence events into a narrow time window. It is the case of an exponential growth population model even if this is not the only model of population growth consistent with the observation of a nearly Poisson mismatch distribution.

When population growth force most of the coalescence events into a narrow time period there is a star–shaped genealogy. A consequence of a star like gene tree is that correlation between coalescence times created by the common history of the individuals are relatively unimportant. Thus information carried by the topology of the tree can be neglected.

In this framework we use information on demographic events contained in the frequency distribution of pairwise coalescence times. In particular, we are concerned with sample of genetic material found at scene of crime for which there is reason to believe it is from the perpetrator of the crime.

A natural question is the following: if the sample found at the scene of crime and the sample of the suspect pertain to the same population or to the same racial group?

To answer this question we propose to use the frequency distribution of pairwise coalescence times as a distribution we name "reference distribution" for the population we conjecture that individuals pertain.

In order to justify this way to proceed we observe that people have a history similar to those of the population from which they are sampled. Another observation is that two people pertaining to the same racial group are in time less distant than two people who pertain to different populations or racial groups.

As showed in the previous section, the distribution of coalescence times of two individuals, two Italians in our simulations, is similar to the reference distribution of the population to whom they pertain, in particular their modes are very closed. Instead the coalescence times distribution of an italian individual and an african one is sensitively different from the reference distribution of italian population.

In spite of the simple models we adopted in this paper and the simple identification problem we formulated, our results suggest that the distribution of pairwise coalescence times contains information suited to forensic identification purposes.

REFERENCES

BALL, R. M. NEIGEL, J. E. and AVISE, C., 1990 Gene genealogies within the organismal pedigrees of random mating populations. Evolution **44:** 360–370.

BATAILLE, M., CRAINIC, K., LETERREUX, M., et al., 1999 Multiplex amplification of mitochondrial DNA for human and species identification in forensic evaluation. Forensic Sci. Intern. **99:** 165–170.

DONNELLY, P., TAVARÈ, S., 1995 Coalescent and genealogical structure under neutrality. Annu. Rev. Genet. **29:** 401–421.

DURRETT, R. 2002 Probability models for DNA sequence evolution. Springer–Verlag, New York.

EWENS, W. J., 1979 Mathematical population genetics. Springer-Verlag, Berlin.

GRIFFITHS, R. C., 1989 Genealogical tree probabilities in the infinitely many site model. J. Math. Biol. **27:** 667–680.

GRIFFITHS, R. C., TAVARÈ, S., 1994 Ancestral inference in population genetics. Statistical Science **9:** 307–319.

LEVIN, B. C., CHENG, H. Y. and REEDER, D. J., 1999 A human mitochondrial DNA standard references material for quality control in forensic identification, medical diagnosis, and mutation detection. Genomics **55:** 135–146.

MILLER, K. W. P., DAWSON, J. L.and HAGELBERG, E., 1996 A concordance of nucleotide substitutions in the first and second hypervariable segments of the human mtDNA control region. International Journal of Legal Medicine **109:**107–113.

NEUHAUSER, C. 2001 Mathematical models in population genetics in *Handbook of Statistical Genetics* edited by D.J. Balding et al. Wiley, Chichester.

NORDBORG, M. Coalescent Theory, 2001 in *Handbook of Statistical Genetics* edited by D.J. Balding et al. Wiley, Chichester.

ROGERS, A. R. and HARPENDING, H., 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol **9:** 552–569.

SLATKIN, M. and HUDSON, R., 1991 Pairwise comparison of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.

SCHNEIDER, S. and EXCOFFIER, L., 1999 Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. Genetics **152:** 1079–1089.

STONEKING, M., SHERRY, S. T., REDD, A. J. and VIGILANT, L., 1992 New approaches to dating suggest a recent age for the human mtDNA ancestor. Phil. Trans. R. Soc. Lond. B. 337.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **123:** 597–601.

TAVARÈ, S., BALDING, D. J., GRIFFITHS, R.C. and DONNELLY, P., 1997 Inferring coalescence times from DNA sequence data. Genetics **145:** 505–518.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theoret. Popul. Biol. **7:** 256–276.

WILSON, I. J., WEALE, M. E., BALDING, D. J., 2000 Inferences from DNA data: population histories, evolutionary processes, and forensic match probabilities. PostScript preprint, Department of Mathematical Sciences, University of Aberdeen.