J. R. Statist. Soc. A (2003)
166, *Part* 2, *pp.* 155–201

Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities

Ian J. Wilson,

University of Aberdeen, UK

Michael E. Weale

University College London, UK

and David J. Balding

Imperial College of Science, Technology and Medicine, London, UK

[*Read before* The Royal Statistical Society *on Wednesday, November 13th, 2002, the President*, Professor P. J. Green, *in the Chair*]

Summary. We develop a flexible class of Metropolis–Hastings algorithms for drawing inferences about population histories and mutation rates from deoxyribonucleic acid (DNA) sequence data. Match probabilities for use in forensic identification are also obtained, which is particularly useful for mitochondrial DNA profiles. Our data augmentation approach, in which the ancestral DNA data are inferred at each node of the genealogical tree, simplifies likelihood calculations and permits a wide class of mutation models to be employed, so that many different types of DNA sequence data can be analysed within our framework. Moreover, simpler likelihood calculations imply greater freedom for generating tree proposals, so that algorithms with good mixing properties can be implemented. We incorporate the effects of demography by means of simple mechanisms for changes in population size and structure, and we estimate the corresponding demographic parameters, but we do not here allow for the effects of either recombination or selection. We illustrate our methods by application to four human DNA data sets, consisting of DNA sequences, short tandem repeat loci, single-nucleotide polymorphism sites and insertion sites. Two of the data sets are drawn from the male-specific Y-chromosome, one from maternally inherited mitochondrial DNA and one from the β-globin locus on chromosome 11.

Keywords: Forensic identification; Human history; Markov chain Monte Carlo methods; Population genetics; Statistical genetics

1. Introduction

Underlying a sample of deoxyribonucleic acid (DNA) sequence data is a complex pattern of dependences that reflects the ancestral relationships between the sequences. In the absence of *recombination*, these relationships can be represented by a genealogical tree for which each tip, or leaf, corresponds to a sequence at the present time. Moving towards the root of the tree corresponds to going backwards in time, and branches merge, or 'coalesce', when the corresponding DNA sequences last had a common ancestor. The root of the tree represents the most recent common ancestor (MRCA) of all the sequences in the sample.

Address for correspondence: Ian J. Wilson, Department of Mathematical Sciences, Meston Building, University of Aberdeen, Aberdeen, AB24 3UE, UK.
E-mail: i.wilson@maths.abdn.ac.uk

 0964–1998/03/166155

Although the underlying genealogical tree is crucial to modelling the dependence structure of a DNA sample, it is effectively ignored by traditional methods of analysing DNA sequence data which, for example, are based on averaging pairwise statistics over all pairs in the sample. In recent years, however, important advances have been made towards the goal of fully likelihood-based statistical inference from population genetics data (Griffiths and Tavaré, 1994; Kuhner *et al.*, 1995, 1998; Beerli and Felsenstein, 1999, 2001; Beaumont, 1999; Anderson *et al.*, 2000; Bahlo and Griffiths, 2000; Stephens and Donnelly, 2000; Chikhi *et al.*, 2001; Donnelly *et al.*, 2001; Nielsen and Wakeley, 2001; Markovtsova *et al.*, 2000a, b). The key developments underpinning these advances involve

(a) an increasingly flexible class of stochastic models for genealogical trees, based on the coalescent model of Kingman (1982), and
(b) computational techniques such as Markov chain Monte Carlo (MCMC) methods and methods based on importance sampling.

However, implementing these models and algorithms remains extremely challenging because of the complexity of the processes underlying the data, which include historical patterns of migration, mating behaviour and population growth, as well as *mutation* and *selection*. Moreover, if recombination cannot be ignored then there may be different genealogical trees for different segments of the sequences (Griffiths and Marjoram, 1997; Kuhner *et al.*, 2000; Nielsen, 2000; Fearnhead and Donnelly, 2002).

Wilson and Balding (1998) developed a Metropolis–Hastings algorithm for completely linked 'microsatellite', or short tandem repeat (STR), loci and reanalysed the human *Y-chromosome* STR data set of Cooper *et al.* (1996), described later in Section 2.1. They found evidence for a relatively small effective population size for human males (point estimates around 3000) and for a short time since the most recent common ancestor, TMRCA, point estimates around 30000–40000 years. However, the modelling assumptions employed by Wilson and Balding (1998) were limited to the standard coalescent (Section 3.1.1) and the stepwise mutation model (Section 3.2.1), and the sensitivity of the final inferences to these modelling assumptions was not investigated.

Here, we extend the analyses of Wilson and Balding (1998) by permitting changes in population size, or structure or both. Inevitably, fully realistic models for the historical patterns of human mating and migration remain outside our grasp. However, we can implement models that generalize those of Wilson and Balding (1998), and hence explore sensitivity to their modelling assumptions, and which capture at least some, and possibly most, of the major underlying demographic effects.

Further, we extend the mutation model of Wilson and Balding (1998) to permit the analysis of a wide range of DNA sequence data for which recombination and selection can be neglected. For the parts of the human *genome* which are subject to recombination, its effects can often be ignored for sequences of up to a few thousand *base pairs* (bp). However, recombination rates appear to be highly variable so some much larger chromosome segments may be unaffected by recombination, and some short segments may be heavily affected (Goldstein, 2001).

Inferences 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, from conservation genetics (Beaumont, 2001) to mapping disease genes (Clayton, 2000). We illustrate this point by elaborating an application of our inferential framework to the assessment of DNA profile evidence. Specifically, we extend our MCMC algorithm by introducing an additional, no-data node, to obtain match probabilities for forensic identification by using mitochondrial DNA miniprofiles.

The present paper is addressed in part to statisticians who may be unfamiliar with some of the genetics terminology. To assist such readers we have included a very brief glossary of terms in Table 1, and each term included is highlighted in italics at first use. Table 1 also includes a list of genetics abbreviations. For further background reading, Hartl and Clark (1997) provide a popular introduction to population genetics, and more advanced material may be found in Balding *et al.* (2001).

2. Data

2.1. Y-chromosome haplotypes

STR *alleles* consist of a short DNA sequence motif, e.g. GATA, multiply repeated (Goldstein and Schlötterer, 1999). Cooper *et al.* (1996) reported the numbers of STRs at each of five *loci* on the non-recombining part of the human Y-chromosome, for 174 apparently unrelated men from East Anglia (UK). A further 23 men from northern Nigeria and 15 from Sardinia were also typed. Although the three groups sampled cannot be regarded as representative of all human populations, they include both African and non-African populations, which is important in view of the prominence of the theory of a recent African origin of modern humans (Relethford, 1998). A further advantage of this data set is that the STR unit is the same at each locus, which makes more plausible an assumption of a common mutation mechanism for all loci.

Table 1. Terminology and abbreviations used in the text†

Term	*Definition*
Allele	Possible state of the DNA sequence (or feature derived from it) at a *locus*
Base pair (bp)	Unit of DNA sequence length, equal to the number of *nucleotides*
Chromosome	Can here be regarded as a long DNA sequence
Genome	Total genetic inheritance of an organism; the human genome consists of 23 *chromosome* pairs (one maternal and one paternal) plus *mitochondrial DNA*
Haplotype	*Alleles* at two or more *loci* on the same *chromosome* (cf. *genotype*: unordered *allele* pairs {maternal, paternal} at one or more *loci*)
Locus (plural loci)	Specified site or short region on a *chromosome*
Mitochondrial DNA	Circular chromosome inherited maternally
MRCA	Most recent common ancestor
Mutation	Process that changes the *allele* at a *locus*
Nucleotide	DNA sequence unit which takes one of four types, denoted A, C, G and T
Polymorphism	*Locus* at which more than one *allele* arises in a given population (cf. *monomorphism*: an invariant *locus*)
Recombination	Exchange of DNA between maternal and paternal *chromosomes* (not *mitochondrial DNA* or *Y*) during the formation of sperm and egg cells
Selection	Process whereby advantageous *alleles* tend to become more common and disadvantageous *alleles* less common (cf. *neutral*: not subject to selection)
SMM	Stepwise mutation model
SNP	Single-nucleotide polymorphism
STR	Short tandem repeat
TMRCA	Time since most recent common ancestor
Transition	Mutation involving a substitution either of a purine *nucleotide* (A or G) by the other or of a pyrimidine (C or T) by the other
Transversion	Substitution of a purine *nucleotide* by a pyrimidine, or vice versa
UEP	Unique event polymorphism
YAP	Y-chromosome Alu polymorphism
Y-chromosome	Sex-determining *chromosome*, borne only by males

†The definitions are not fully general but are intended as a guide.

In addition to the STR data, Cooper *et al.* (1996) reported for each Y-chromosome the presence or absence of the so-called *Alu* insertion sequence at a particular locus, known as the *Y-chromosome Alu polymorphism* (YAP) locus described in Hammer (1994). In formulating likelihoods for the YAP data, we face an ascertainment problem which is common to many DNA sequence data sets. Cooper *et al.* (1996) chose to type this locus because it was known to be polymorphic in many human populations (and many other potential loci that were not known to be polymorphic were not typed). Inferences which can be drawn under these circumstances will differ in general from inferences which would be justified if the locus had been chosen to be typed 'at random'.

The human Y-chromosome data recorded by Ruiz Linares *et al.* (1996) consist of five *dinucleotide* STR loci (including one monomorphic locus), a tetranucleotide STR (one of those included in the study of Cooper *et al.* (1996)), the YAP locus and a single-nucleotide polymorphism (SNP) site. Data were obtained from 13 worldwide populations, although the sample sizes are often small (see Table 5 later). Ruiz Linares *et al.* (1996) noted substantial geographic clustering of the observed *haplotypes* and inferred that the MRCA of human Y-chromosomes cannot have been very recent, but they did not give an estimate of this time.

2.2. β-globin sequences

Harding *et al.* (1997) analysed a subset of the data of Fullerton *et al.* (1994), consisting of 61 DNA sequences from the Melanesian population of Vanuatu. The 3000 bp region sequenced encompasses the β-globin gene. One end of this region was later identified as a recombination hot spot and 330 bp at that end were ignored to permit analyses based on an assumption of no recombination.

Harding *et al.* (1997) adopted the 'infinite sites' mutation model, which implies that at any one DNA site there has been no more than one mutation since the MRCA of the sample. The full data set was not consistent with the assumption, but Harding *et al.* (1997) discarded four sequences which seemed to have been affected by recombination, resulting in a 57-sequence data set that was consistent with infinite sites. In contrast with the geographic clustering of Y-chromosomes reported by Ruiz Linares *et al.* (1996), Harding *et al.* (1997) reported substantial haplotype diversity in the Vanuatu sample, with all known haplogroups (groups of similar haplotypes) represented in this sample from a single isolated location. Similarly, the estimate of Harding *et al.* (1997) of the effective size of the ancestral Vanuatu population is approximately the same as many estimates of the effective size of the entire human population, suggesting that the current geographic isolation is unimportant in explaining the observed genetic diversity. They obtained a point estimate of 895000 years for the TMRCA of the Vanuatu β-globin sequences, which is longer than corresponding estimates based on worldwide Y-chromosome data, even when the fourfold larger population size of β-globin sequences is taken into account. (Each child receives two β-globin sequences from its parents, but on average only half a Y-chromosome. Under simple models, the expected TMRCA is proportional to the population size and so would be fourfold higher for β-globin sequences than for Y-chromosomes).

The infinite sites assumption adopted by Harding *et al.* (1997) is crucial to some methods of analysing DNA sequence data, since it implies that all mutations which have occurred are directly visible in the data; none have been 'overwritten' by a subsequent mutation. The assumption is not valid for many data sets. It is neither required nor advantageous under the framework that is developed here, but it can readily be implemented and we do so later (Section 5.3) to permit a comparison with the results of Harding *et al.* (1997).

2.3. Mitochondrial minisequences

The maternally inherited mitochondrial DNA has been widely used to infer aspects of human female population histories (Cann *et al.*, 1987; Sykes, 1999). Because it exists in multiple copies throughout each cell, compared with just one copy of nuclear DNA from each parent, mitochondrial DNA is easier to type from small and/or degraded samples. In recent years Neanderthal and ancient Australian mitochondrial DNA have been successfully typed (Krings *et al.*, 1997; Ovchinnikov *et al.*, 2000; Adcock *et al.*, 2001).

Mitochondrial DNA typing is also useful in forensic identification (Tully *et al.*, 1996, 2001; Bataille *et al.*, 1999), mostly for samples of shed hair, which are often recovered at crime scenes and which contain little or no nuclear DNA. In addition, mitochondrial DNA has been successfully typed from DNA poor sources such as saliva on stamps (Allen *et al.*, 1998) and tooth root dentine (Pfeiffer *et al.*, 1998). However, mitochondrial DNA suffers from a serious drawback in forensic settings: because of the absence of recombination, the mitochondrial DNA sequences of two apparently unrelated individuals can be identical, or very similar, owing to shared inheritance from a common maternal ancestor, possibly many generations in the past. This makes it difficult to assess the evidential weight of matching mitochondrial DNA profiles in a way which is fair but makes efficient use of the data.

The UK Forensic Science Service (FSS) employs mitochondrial DNA 'minisequences' consisting of 12 loci in the mitochondrial DNA control region that display a high level of genetic variation, while being relatively quick to type. The 12 loci comprise 10 SNP sites, a dinucleotide STR locus and a locus made up of multiple copies of the C-base (a poly-C region) (Tully *et al.*, 1996). The poly-C locus was excluded from our analyses, both because there is little available information from which to formulate an appropriate mutation model and because it is not always utilized in forensic case-work owing to its high rate of within-individual variation. We analyse an FSS database of 297 minisequences, obtained from apparently unrelated UK residents, 152 with primarily European ancestry, 103 of Afro-Caribbean origin and 42 with Asian ancestry.

The ascertainment bias problem discussed in Section 2.1 arises again for mitochondrial DNA minisequences: the 12 loci were selected in part because preliminary studies indicated high variability at these loci.

3. Models

3.1. Demography

3.1.1. Standard coalescent

The coalescent is a stochastic model for the genealogical tree representing the ancestral relationships between a sample of n DNA sequences. The sequences are regarded as labelled, to avoid combinatorial complications, but they are not yet observed. The model has two attractive features: it is mathematically tractable, and it approximates the distribution of genealogical trees under an important class of *neutral* population genetics models, including the Wright–Fisher model of a random-mating population of constant size N, and the general exchangeable models of Cannings (1974). To recover these approximations, 1 unit of 'coalescent' time must be interpreted as N/σ^2 generations, where σ^2 denotes the variance in the number of 'offspring' of a sequence in the next generation. We assign $\sigma^2 = 1$ here but note that the mating behaviour of men differs from that of women and σ^2 for Y-chromosome sequences may be larger than the value that is appropriate for mitochondrial DNA sequences, and may possibly be much larger than 1.

Note that we use N for the number of chromosomes, and not the number of individuals: for the β-globin data, N sequences correspond to $N/2$ individuals; for Y-chromosome and mitochondrial DNA data, N sequences correspond to $2N$ individuals, N males and N females.

Coalescent time runs backwards, with time $t_0 \equiv 0$ denoting the present, corresponding to the leaves of the tree, whereas t_j, $j \in \{1, 2, \ldots, n-1\}$, denotes the time of the jth most recent coalescence event. In particular, t_{n-1} denotes the time of the root, or MRCA. Under the standard coalescent model, the between-coalescence intervals $t_j - t_{j-1}$ have independent exponential distributions:

$$P(t_j > t | t_{j-1} = t') = \exp\{\beta_j(t' - t)\} \qquad \beta_j \equiv \binom{n+1-j}{2} \tag{1}$$

for $t > t'$. At each coalescence event, all pairs of extant lineages are equally likely to be the pair that coalesces.

Mutations in the standard coalescent model occur along the branches of the tree at the points of a homogeneous Poisson process with rate $\theta/2$. Because of the coalescent time rescaling, this corresponds to a mutation rate of $\mu \equiv \theta/2N$ per locus per generation in a population of N sequences.

Two notable features of the standard coalescent model are

(a) the long period of time (on average, more than half TMRCA) in which the tree has just two lineages and
(b) the high variance in total tree height (the standard deviation (SD) is typically about 60% of the mean).

See Nordborg (2001) for a more detailed introduction to coalescent models.

3.1.2. *Coalescent with population growth*

The standard coalescent model is the special case $\lambda(s) \equiv 1$ of the model in which equation (1) is replaced with

$$P(t_j > t | t_{j-1} = t') = \exp[\beta_j\{\Lambda(t') - \Lambda(t)\}], \tag{2}$$

where $\Lambda(t)$ is an increasing differentiable function with $\Lambda(t) \equiv \int_0^t \mathrm{d}s/\lambda(s)$. The model thus defined approximates the genealogy of a sample drawn from a random-mating population of size $N\,\lambda(t)$ at time $N \int_0^t \lambda(s)\,\mathrm{d}s$ generations ago (Hudson, 1991; Donnelly and Tavaré, 1995). Intuitively, an increment of coalescent time corresponds to more generations when the population size is large than when it is small.

One simple model for a change in population size is the 'k-size coalescent', which in the case $k = 2$ is specified by

$$\begin{aligned} \lambda(t) &= \alpha, \quad \Lambda(t) = t/\alpha & 0 < t < t_{\mathrm{g}}, \\ \lambda(t) &= 1, \quad \Lambda(t) = t_{\mathrm{g}}/\alpha + t - t_{\mathrm{g}} & t > t_{\mathrm{g}}, \end{aligned} \tag{3}$$

corresponding to a population of constant size N until time t_{g}, after which it instantly attains its present size, $N\alpha$.

Pure exponential growth at rate r per generation corresponds to

$$\begin{aligned} \lambda(t) &= c\exp(-Rt), \\ \Lambda(t) &= c\exp(Rt)/R, \end{aligned}$$

where $R \equiv Nr$ and c is an arbitrary constant. When $R > 0$, there are fewer recent coalescences than under the standard coalescent model (with the same expected total branch length, which can be achieved by an appropriate choice of c) and hence more mutations represented only once in the data. When $R < 0$ each coalescence event has positive probability that it does not occur in finite time, leading to clusters of sequences separated by an infinity of mutations.

Pure exponential growth is unlikely to provide a good model for global human population size, since recent high growth rates would imply a vanishingly small population size a few thousand years ago. Marjoram and Donnelly (1994) considered a two-parameter model for which

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

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

$$N_c \equiv N(1+r)^{Nt_g} \approx N\exp(Rt_g).$$

We adopt this model below, and for convenience we refer to it as the 'coalescent with growth', even though other formulations of population growth are possible. Under this model,

$$\Lambda(t) = \begin{cases} \exp(-Rt_g)\{\exp(Rt) - 1\}/R & 0 < t < t_g, \\ t - t_g + \{1 - \exp(-Rt_g)\}/R & t > t_g, \end{cases}$$

and the coalescence time distributions (2) become

$$P(t_j > t | t_{j-1} = t') = \begin{cases} \exp[\beta_j\{\exp(Rt') - \exp(Rt)\}\exp(-Rt_g)/R] & t' < t < t_g, \\ \exp\{\beta_j(t_g - t + [\exp\{R(t' - t_g)\} - 1]/R)\} & t' < t_g < t, \\ \exp\{\beta_j(t' - t)\} & t_g < t' < t. \end{cases} \quad (4)$$

The coalescent with growth model reduces to the standard coalescent model both when $t_g = 0$ and in the limit as $R \to 0$. In the examples below we rely on background information to justify an *a priori* assumption of $R \geqslant 0$ but note that $R < 0$ in our model does not lead to infinite coalescence times, as is the case for pure exponential growth.

3.1.3. Coalescent with population splitting

Human populations are often subdivided, in some cases by cultural barriers, but most obviously by geographical barriers or distance. The two Y-chromosome samples that were described in Section 2.1 are both divided into subsamples obtained from geographically distinct human subpopulations such that individuals are more likely to mate within their subpopulation than outside it. Modelling population subdivision may be crucial to the interpretation of the two data sets, but this is not necessarily the case: a relatively low level of migration can suffice largely to eliminate the effects of subdivision (Hartl and Clark, 1997), and there is evidence for large scale migrations throughout human history and prehistory (Cavalli-Sforza and Cavalli-Sforza, 1995).

One popular approach to modelling subdivision is based on the island model of Wright (1931). This is an equilibrium model in which each pair of subpopulations exchanges migrants at points of a homogeneous Poisson process of given rate; see Wilkinson-Herbots (1998) for further details and Bahlo and Griffiths (2000) and Beerli and Felsenstein (2001) for coalescent-based inference under this model. However, the equilibrium assumption underlying the island model is questionable for human populations. In addition, the number of migration parameters grows with the square of the number of populations, becoming unmanageably large for more than a handful of populations, and an assumption of a common migration rate is usually highly unrealistic.

A simple non-equilibrium model which allows changes in population structure is that em-

ployed by Weir and Cockerham (1984), which posits a random-mating ancestral population of size N until Nt_a generations ago, when it split into isolated subpopulations of equal sizes. We adopt this model with two extensions, to allow

(a) the ith subpopulation to have size $N\alpha_i$, with $\Sigma\,\alpha_i = 1$ and
(b) population bifurcations occurring at different times.

The process of subpopulation splits creates a population 'supertree', with one leaf for each subpopulation, which we model separately from the underlying genealogical tree. Fig. 1 illustrates a realization of a genealogy in four subpopulations under this 'splitting' model.

For the coalescent approximation to the splitting model, the genealogical tree underlying the sample from each subpopulation is given by the k-size coalescent, introduced above at expression (3). This 'coalescent with splitting' rules out certain coalescence events (between sequences in different subpopulations) which would be permitted under the standard coalescent model. However, those coalescences which are permitted occur at a higher rate because of the smaller

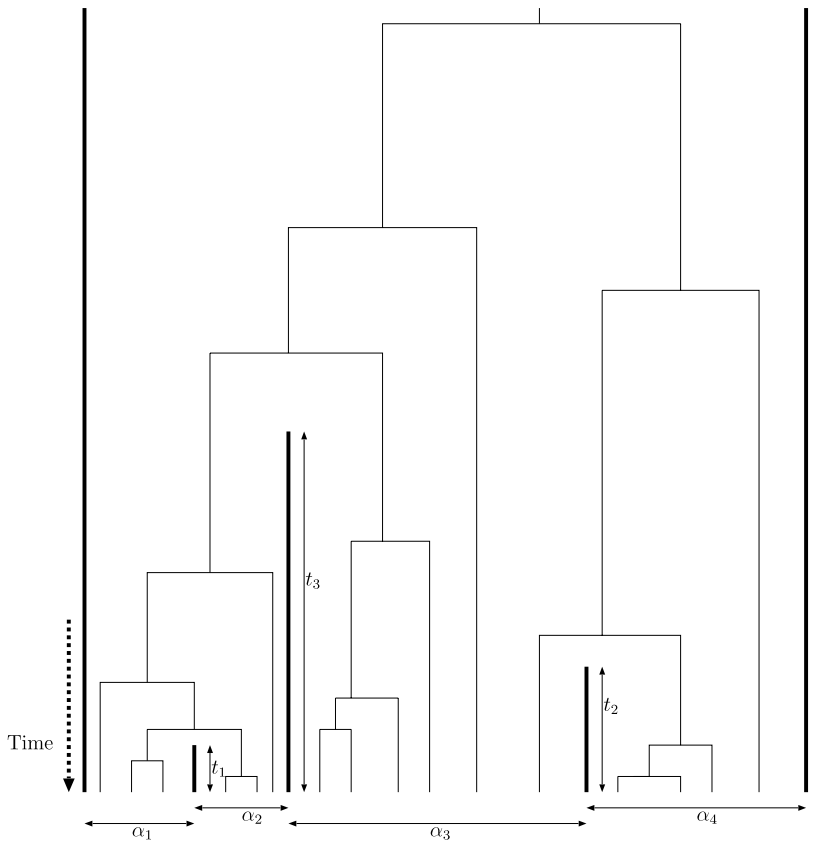

Fig. 1. Genealogy under the 'splitting' model of population subdivision: the single ancestral subpopulation split t_3 ($\equiv t_a$) coalescent time units ago into two subpopulations, each of which subsequently split to result in four current subpopulations, whose sizes form proportions α_i, $i = 1,\ldots,4$, of the total population size N; the subpopulation sample sizes are 3, 3, 6 and 4, corresponding to the number of leaves (terminal nodes) (the broken arrow in the left-hand margin indicates the direction of true time; coalescent time runs in the reverse direction); the figure has been contrived to avoid lineages crossing each other, but the subpopulations are not spatially ordered under our model and such crossings would normally be needed

subpopulation sizes. Overall, the rate of coalescences can be either greater or less than under the standard coalescent model. However, if the subpopulation sizes are approximately equal, and the subsample sizes are also approximately equal, then the overall rate of coalescences is reduced by population splitting, and the expected TMRCA is increased.

The coalescent-with-splitting model of population structure remains unrealistic for many human populations, because it disallows the exchange of migrants between subpopulations after they split. Nielsen and Wakeley (2001) formulated a model which incorporates migration after a split, but their implementation is limited to two subpopulations, without population growth. Our simpler model captures much of the effect of population subdivision with relatively few additional parameters.

3.1.4. Coalescent with splitting and growth

In the analyses below, we implement a model which incorporates both the coalescent with growth (Section 3.1.2) and the coalescent with splitting (Section 3.1.3). A genealogy under this model is illustrated in Fig. 2. We restrict attention to the case of a common growth rate in all subpopulations and at all times after the start-of-growth time t_g. The extension to growth rates which change after each split is straightforward in principle, but the data are only weakly informative about growth rates and so estimation may be poor.

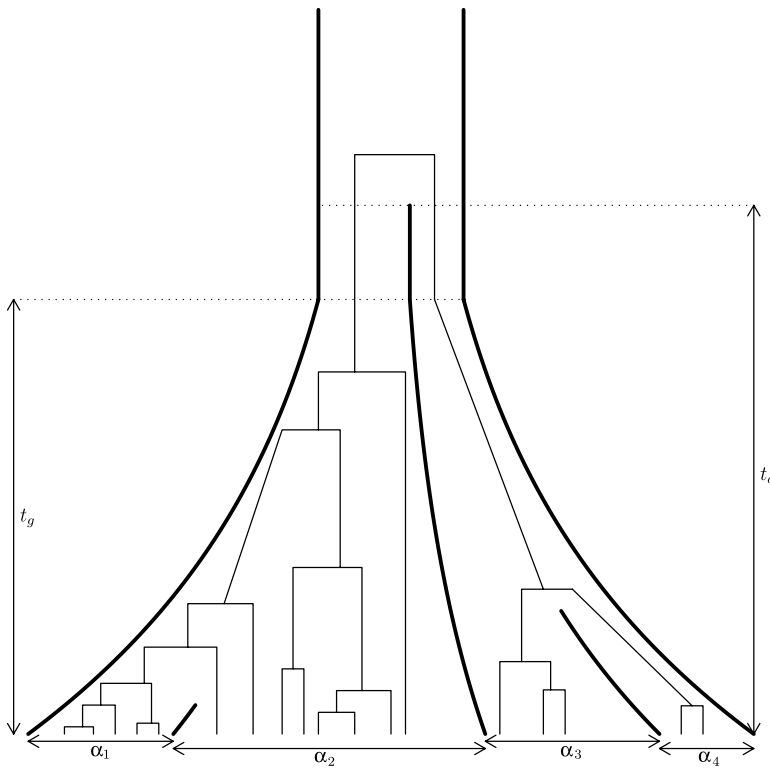

Fig. 2. Genealogy under the 'splitting-with-growth' model: the first subpopulation split occurs at time t_a (the horizontal axis represents population size; after time t_g the total population size grows exponentially; other details are the same as for Fig. 1)

3.2. Mutation

3.2.1. Short tandem repeat loci

Since different STR loci are usually widely separated, it seems reasonable to assume independence of the mutation processes at distinct loci. The most widely adopted mutation model for an STR locus is the stepwise mutation model (SMM) in which the mutant allele differs from its parent by one repeat unit. Steps in each direction are equally likely, irrespective of the current allele length. Although there is evidence (Brinkmann *et al.*, 1998; Kayser *et al.*, 2000) for the SMM being close to the actual mutation process at STR loci, there is also evidence of deviations from the model, which we now briefly discuss.

Cooper *et al.* (1999) analysed Y-chromosome STR data by using coalescent models and reported evidence for a mutation bias, with mutations leading to increases in allele length more often than decreases. However, the signal in the data for such a bias lies in the skewness of the allele frequency distribution, and this may have other causes. The direct observations of Kayser *et al.* (2000) do show a non-significant excess of increases (10) over decreases (4), but Brinkmann *et al.* (1998) reported a slight excess in the other direction.

There is direct evidence for the strict one-step model being false: Kayser *et al.* (2000) observed one mutation (out of 14) which altered the allele length by two repeat units and Brinkmann *et al.* (1998) reported one two-step mutation and 22 one-step mutations. Pritchard *et al.* (1999) fitted a model to Y-chromosome STRs in which the change in the number of repeat units forms a geometric(p) random variable. They estimated p to be close to 1, in which case their model is difficult to distinguish from the SMM with a slightly higher mutation rate.

Both Kayser *et al.* (2000) and Brinkmann *et al.* (1998) reported an apparent correlation between allele length and mutation rate. In the former study the correlation was weak, and in the latter case an important contribution to the correlation seems to have arisen from an absence of mutations observed at loci with mean allele length under 10 repeats. No such loci are included in the two Y-chromosome data sets that we analyse below.

Extensions to the SMM to allow for mutation rate increasing with length, mutation bias or multistep mutation events can readily be implemented within our framework. For the reasons indicated above, and to keep the presentation as simple as reasonably possible, we choose not to implement any such extensions here and retain the SMM adopted by Wilson and Balding (1998).

The SMM has no equilibrium distribution, and so there is no natural prior distribution for the STR repeat number at the root of the genealogical tree. A prior which is uniform on the positive integers, although improper, leads to a proper posterior distribution and is adopted below. Although the SMM does not constrain allele lengths to be positive, we find that, conditional on the current allele sizes, the probability that an ancestral allele is assigned a non-positive length is negligible in practice.

3.2.2. Single-nucleotide polymorphism sites: recurrent mutation

We assume that insertions, deletions and translocations of nucleotides are sufficiently rare that they can be ignored for the timescales that are relevant here. Mutations thus consist only of the substitution of one nucleotide by another. The mitochondrial DNA minisequence sites are separated by many nucleotides, and so it is natural to assume that the substitutions at distinct sites are mutually independent. An SNP mutation model can then be specified by a continuous time Markov chain on the states A, C, G and T.

The simplest such model is the Jukes–Cantor model (Jukes and Cantor, 1969), in which all possible substitutions are equally likely, so that the chain has a uniform stationary distribution:

$\pi_A = \pi_C = \pi_G = \pi_T = \frac{1}{4}$. Felsenstein (1981) introduced the F81 model, with an arbitrary stationary distribution: substitutions occur at rate β and the mutant nucleotide is A, C, G or T with probabilities π_A, π_C, π_G and π_T. Note that a nucleotide may be substituted for one of the same type, so that the effective mutation rate is less than β.

We adopt here the F84 model, an extension of the F81 model that is similar to the model of Hasegawa *et al.* (1985). It has been implemented since 1984 in the program DNAML in the PHYLIP suite of programs (described in Felsenstein and Churchill (1996)). Under the F84 model, a parameter α specifies the nominal rate of additional mutations restricted to be *transitions*. If the current nucleotide is either of the purines (A or G), then the mutant nucleotide is A or G with probabilities $\pi_A/(\pi_A + \pi_G)$ and $\pi_G/(\pi_A + \pi_G)$, and similarly for the pyrimidines. The stationary distribution is unaffected by the additional transitions, and the overall effective mutation rate is

$$\mu_{SNP} = \beta(1 - \pi_A^2 - \pi_G^2 - \pi_C^2 - \pi_T^2) + 2\alpha\left(\frac{\pi_A\pi_G}{\pi_A + \pi_G} + \frac{\pi_C\pi_T}{\pi_C + \pi_T}\right) \tag{5}$$

per nucleotide per generation.

3.2.3. Unique event polymorphism loci

If $\theta\ (\equiv 2N\mu) \ll 1$ at a biallelic locus, it may be reasonable to assume that there was only one mutation event underlying the observed polymorphism. Although this unique event polymorphism (UEP) assumption is not required in our framework, if valid it can greatly reduce the size of the tree space that must be explored, allowing both computational efficiency and more precise inferences. These advantages are enhanced if it is known which of the two alleles is ancestral, in which case any two haplotypes with the non-ancestral state must be more closely related than two haplotypes with differing states.

The YAP (Section 2.1) was assumed by Cooper *et al.* (1996) to be a UEP. The Alu element is widely interspersed in the human genome and seems capable of being inserted at any location (Sherry *et al.*, 1997), so two inserts at precisely the same location seem very unlikely.

For SNPs, many researchers assume the 'infinite sites' mutation model (Section 2.2) under which all SNPs are also UEPs. Since per site θ for humans is on average of the order of 10^{-3}, this assumption seems reasonable provided that mutation is reasonably homogeneous over sites. In contrast with the YAP, for which the UEP assumption implies that the MRCA sequence must lack the Alu insert, the ancestral SNP state is usually unknown, though it is sometimes assumed to be the most common observed state.

4. Methods

4.1. Coalescent models as Bayesian priors

Coalescent models have revolutionized population genetics in the past decade by moving the emphasis away from prospective modelling of populations, which allows at best only crude comparisons with observed data, and towards a retrospective description of the genealogy of a sample. Although coalescent-based methods permit the specification of a likelihood function, many of the new inferential methods are not fully likelihood based, primarily because of the computational complexity (see Fu and Li (1999) for a review). Likelihood-based methods have been formulated in recent years, either via MCMC or via importance sampling approaches (Stephens, 2001), thereby bringing to population genetics problems the benefits of statistical efficiency and quantitative model comparison.

Because of the complex models and substantial background information, the Bayesian paradigm provides an appropriate framework for statistical inference in the present setting. The demographic models that were introduced in Section 3.1 specify prior distributions for the genealogical tree underlying a sample of DNA sequences, and inference about aspects of this tree proceeds via its posterior distribution given the observed data (and the priors for the mutation rate and other parameters).

The development of MCMC algorithms to approximate the required posterior distribution is challenging for several reasons, including the complexity of the likelihood computations. Wilson and Balding (1998) developed a Metropolis–Hastings algorithm, based on an augmented data approach in which the allelic states at all internal nodes of the tree (i.e. at each coalescence) were regarded as auxiliary variables. The likelihood calculations were thereby greatly simplified, at the cost of a larger parameter space. Thanks to the simpler likelihood calculations, a wide class of proposal distributions for exploring the space of genealogical trees was available, allowing an algorithm with good mixing properties to be developed. We retain these features in the algorithms that are presented here, together with additional features to incorporate the demographic and mutation models introduced earlier.

4.2. Algorithm

4.2.1. *The BATWING software*

A Metropolis–Hastings algorithm to investigate the models and data sets described above has been implemented in a C program, for which we have adopted the acronym BATWING (Bayesian analysis of trees with internal node generation). UNIX, Windows and Macintosh versions are freely available at

`http://www.maths.abdn.ac.uk/~ijw/`

Also available is the *BATWING User Guide* which explains how to use the program.

4.2.2. *Proposal distributions*

BATWING's proposal distribution for updating the tree topology is similar to that of MICSAT, described in Wilson and Balding (1998) and also available at the above Web site. Briefly, candidate trees are obtained by selecting an internal node at random on the current tree and choosing a new location at random, but locations near nodes of similar allelic type are more likely to be chosen than locations near dissimilar nodes. For demographic models involving splitting, and for mutation models involving UEPs, the obvious modifications are implemented: the initial state and any proposed state are disallowed if they contravene the model. Thus, for example, if the descendants of the selected node are from two different subpopulations, then the node can be relocated only before the time at which those two subpopulations split.

In addition to updates which change the topology of the tree, BATWING also proposes updates to coalescence times (keeping internal haplotypes constant) and updates to internal haplotypes (keeping coalescence times constant). For a UEP site with unknown ancestral state, proposals are made to change the root UEP haplotype at random among those consistent with the tree.

The method for updating the population supertree topology and splitting times is based on a representation of trees given in Mau *et al.* (1999). A planar representation of the supertree is made by randomly allocating 'left' and 'right' subpopulations at each splitting event. The tree can then be written as a list of populations, with a unique splitting time associated with each pair of neighbouring populations. Updates of these splitting times are made by choosing a split

at random, and then choosing a new time for this split, uniformly between 0 and the most recent coalescence between two sequences, one from each of the populations. The updated planar representation then implies a new supertree, which may differ in topology as well as splitting times from the current supertree.

Updates to the proportions into which the ancestral population splits are made independently of the supertree updates. Two populations i and j are chosen at random, and α_i', the new proportion in population i, is drawn uniformly in $(0, \alpha_i + \alpha_j)$. Finally, $\alpha_j' = \alpha_j + \alpha_i - \alpha_i'$.

The demographic parameters (N, N_c, r) and μ are each updated on a logarithmic scale by independent uniform perturbations centred at the current value.

4.2.3. Convergence and mixing

Our principal diagnostic tool has been to compare the results from repeat runs of BATWING with widely spaced initial trees, while all other inputs remain unchanged. For this, BATWING includes a 'badness' parameter b_s which can be set in the interval (0,1), with 0 corresponding to the tree obtained by applying a parsimony heuristic to the data, which is expected to produce a plausible, or 'good', tree, whereas 1 corresponds to a random tree under the prior model, which is almost always 'bad' for any given data set. With the badness parameter between 0 and 1 successive coalescences in the starting tree are drawn with probability proportional to $1/(10b_s + d)$ where d is the minimum number of mutations required under our modelling assumptions to obtain one haplotype from the other. The allelic types at the new node are drawn from a discretized Gaussian distribution with SD equal to $d + 10b_s/4$.

In addition, we routinely inspected plots of the log-likelihood and parameter values over each run of BATWING, as well as autocorrelation plots, for signs of non-convergence. Fig. 3 shows trace and autocorrelation function plots for two parameters and the log-posterior density from one of the analyses in the simulation study (Section 5.1), thinned by retaining every fourth output. The plots suggest approximate stationarity, the trace plots showing no obvious trend or jump, and each autocorrelation function decreasing to 0. Mixing seems good for μ but slow for L, which is unsurprising because L is a function of all the branch lengths, and it depends on the population splitting times and the growth parameters. In most studies L is not a parameter of direct interest.

An 'effective' size m_e of a sample of m observations of a stationary time series can be defined as

$$m_e = m/\left(1 + 2\sum \rho_i^2\right), \tag{6}$$

where ρ_i is the lag i autocorrelation (Liu, 2002). For the data of Fig. 3, $m = 5000$ and truncating at lag 50 we estimate $\hat{m}_e = 4600$ for μ and $\hat{m}_e = 580$ for L, which is the smallest $\hat{m}_e$ among the 13 parameters monitored. The only other parameter with $\hat{m}_e < 2000$ is the start-of-growth time t_g, which is strongly correlated with the growth rate r and the population size.

For the data analyses described below, some parameters such as L, r and t_g did sometimes mix slowly, particularly for C96 (Cooper *et al.*, 1996), our largest data set with more than 200 haplotypes. However, over the several years that the present paper has been in gestation, the authors have not detected any problem with convergence or mixing that was not easily diagnosed with the simple tools illustrated above and addressed by lengthening the run. In particular, we have not encountered evidence of multimodality in our models. Although our models are complex and our parameter spaces are large, we may benefit in one sense from the poor information content of the data for the parameters of interest: posterior distributions tend to be diffuse and lacking the 'peaks' and 'troughs' that can lead to pseudoequilibria.

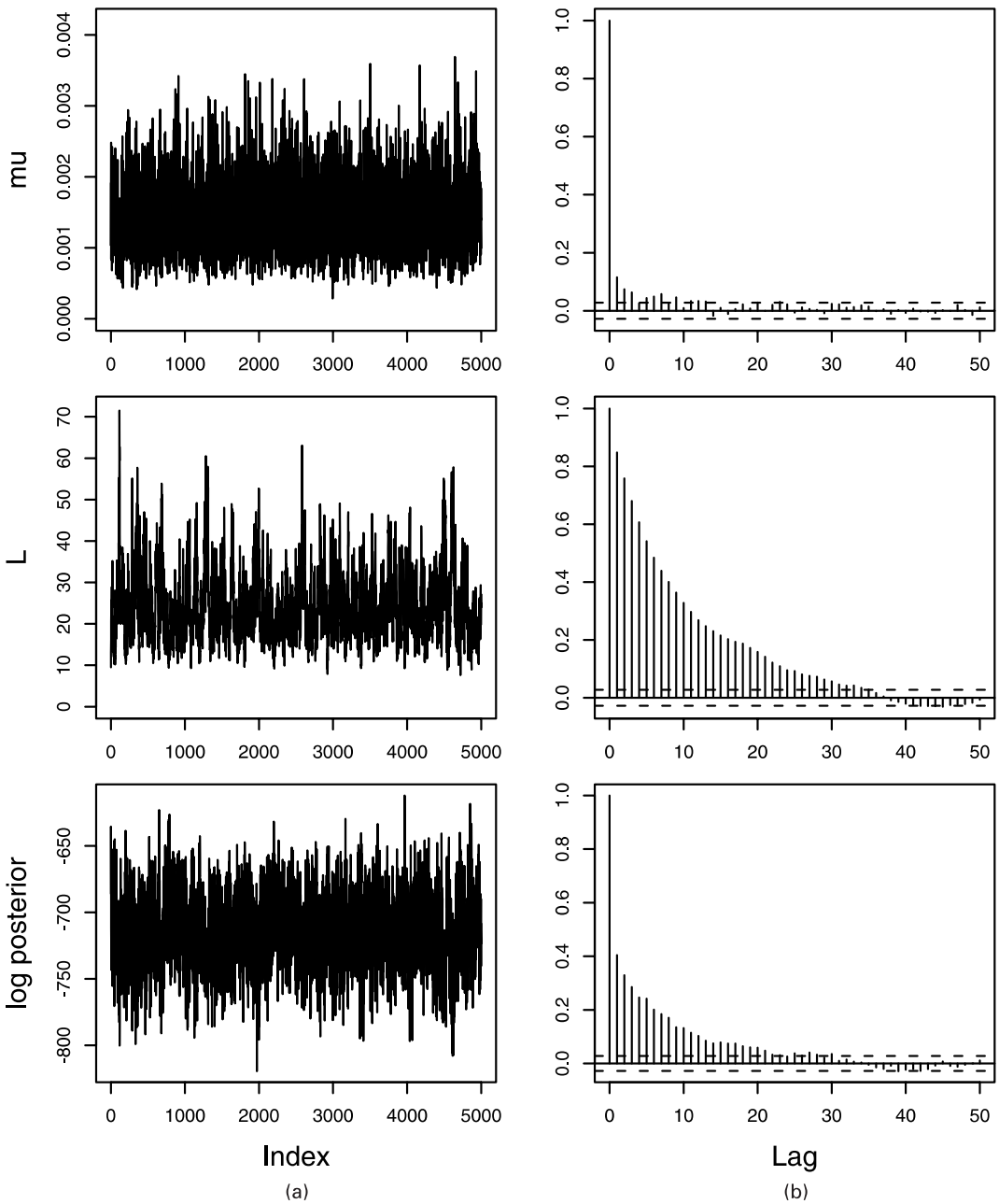

Fig. 3. Diagnostic plots for the BATWING output from one run of the simulation study (a burn-in of 1000 outputs was discarded, and the subsequent outputs were thinned by retaining every fourth output): (a) trace plots and (b) autocorrelations up to lag 50, of the mutation rate μ, the total tree length L and the (unnormalized) log-posterior density

4.2.4. Validation

Irreducibility of the Markov chain implied by the proposal distributions described above is easy to establish when there are no UEP sites with unknown root state. Successive 'cut-and-join' moves can move from any tree to any other. Movement between any two population supertrees can similarly be shown by first changing the underlying genealogical tree, and then successively rearranging splits to obtain the required supertree. With UEPs the cut-and-join moves allow communication between any states that have the same root haplotype. Proposals to change the root UEP haplotype then ensure movement between any two states that are consistent with the data.

In the context of the complex models and data described here, a rigorous verification that the algorithm described above has been correctly implemented in the BATWING software, and usually converges, is not achievable. Some theoretical support is given by Aldous (2000) who showed that, in a no-data setting, a branch swapping algorithm similar to that implemented in BATWING gives convergence in a total variation sense to the appropriate uniform distribution on tree space. The number of steps required is of the order of between n^2 and n^3 moves (Aldous conjectured that n^2 is correct), which is slow compared with the $n \log(n)$ moves that are required for a pack of n cards, randomizing one card at a time.

The authors are also encouraged that results from MICSAT published in Wilson and Balding (1998) have since been verified by using an independent algorithm (Stephens and Donnelly, 2000). Further support comes from a simulation study described later in Section 5.1.

4.3. Mitochondrial DNA match probabilities

Suppose that a mitochondrial DNA sequence is recovered from a sample from the crime scene and found to match the mitochondrial DNA sequence obtained from s, a suspect. This observation supports the hypothesis that s is the source of the crime sample. To assess the strength of this support, we wish to evaluate the probability that x would also match the mitochondrial DNA sequence from the crime scene, where x is an alternative possible culprit whose mitochondrial DNA sequence is unknown. For a discussion of the role of match probabilities in the assessment of forensic identification evidence, see Balding and Donnelly (1995).

If we knew the mitochondrial DNA sequence of any of the maternal ancestors of x, as well as the number of generations separating that ancestor from x, then the probability distribution for the mitochondrial DNA sequence of x could readily be calculated under a given mutation model. Of course, this information is not usually available. However, in forensic case-work a reference sample is usually available consisting of the mitochondrial DNA sequences of n apparently unrelated individuals drawn from the same racial group as x. The $n + 1$ observed mitochondrial DNA sequences, those of s and the reference sample, provide information about ancestral mitochondrial DNA sequences, and hence about the unobserved mitochondrial DNA sequence of x.

Here, we exploit this information via a modification of the algorithm described in Section 4.2. In addition to the genealogical tree underlying the $n + 1$ observed sequences, we introduce a branch connecting the unobserved DNA sequence of x with the tree, writing z for the new node thus introduced (Fig. 4). The additional branch, and the mitochondrial DNA state at z, are updated in the same way as for the other branches, except that, since no data are available for the mitochondrial DNA sequence of x, there is no contribution to the likelihood from the branch connecting z with x.

At each iteration of the modified algorithm, the probability that the mitochondrial DNA sequence of x matches that of s, conditional on the location and state of z, is readily calculated.

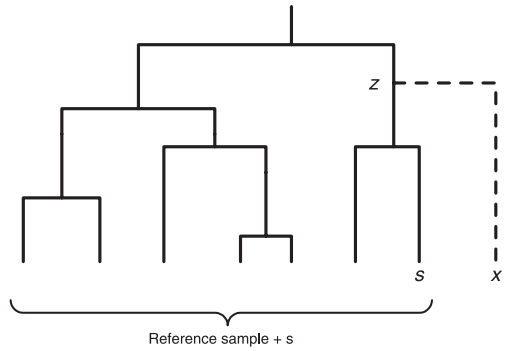

Fig. 4. Representation of the maternal genealogy of $n = 6$ individuals in an anonymous reference database, together with the suspect s and a further individual x regarded as an alternative suspect of unknown mitochondrial DNA type: the node labelled z corresponds to the most recent woman ancestral to both x and at least one of the other $n + 1$ individuals

The average of these conditional probabilities approximates the match probability, given the observations and the standard coalescent model.

5. Analyses

5.1. Simulation studies

A large simulation study was undertaken to check the accuracy of the posterior approximations obtained using BATWING. 100 data sets were simulated from the most complex of our models, the coalescent with splitting and growth (Section 3.1.4). We used three populations, with sample sizes 23, 22 and 15, and the data types were the same as for the C96 data set (five STR loci and one UEP). The SMM (Section 3.2) was employed to generate the STR data and the location of the UEP mutation was chosen at random in the tree. The parameters underlying each simulation (subpopulation sizes, splitting times, start-of-growth time, mutation and growth rates) were obtained via independent draws from the prior distributions that were used to analyse the C96 data set. The model assumptions and prior distributions supplied to BATWING were the same as those used to generate the data.

The average over simulations of $\hat{m}_e$, the estimated effective sample size defined at equation (6), is shown in the first row of Table 2.

For a specified parameter, let H_x indicate whether the $100p\%$ posterior interval, given data x, includes the correct value. If the BATWING program is valid and the runs are sufficiently long then the 'hit rate' $\bar{H}$, the observed average of H_x over the data sets, forms a binomial$(100, p)$ proportion. Hit rates (expressed as percentages) for six parameters are shown in Table 2. In most cases $\bar{H}$ understates the nominal coverage p, but the difference between $\bar{H}$ and p exceeds three SDs only for the growth rate r when $p \geqslant 70\%$.

Although $\bar{H}$ is valuable for ease of interpretation, Rubin and Schenker (1986) pointed to more precise methods of assessing interval coverage. We apply their 'average probability coverage' method to $100p\%$ equal-tailed prior intervals. Given a parameter, say ϕ, write C_x for the posterior coverage, given data x, of an interval (l, u) with prior probability p:

$$C_x \equiv \int_l^u P(\phi|x)\,\mathrm{d}\phi = \int_l^u P(x,\phi)\,\mathrm{d}\phi \Big/ \int\!\!\int P(x,\phi')\,\mathrm{d}\phi'.$$

Table 2. Results for six parameters from a simulation study consisting of 100 data sets generated from the coalescent with splitting and growth model†

p (%)	*Results for the following parameters:*					
	N	μ	T	L	r	t_a
	Average estimated effective sample size					
	2800	15000	5100	1300	6000	5400
	Hit rate $\bar{H}$ (%)					
10	6 (3.0)	12 (3.0)	3 (3.0)	8 (3.0)	12 (3.0)	9 (3.0)
30	24 (4.6)	25 (4.6)	25 (4.6)	30 (4.6)	30 (4.6)	30 (4.6)
50	46 (5.0)	49 (5.0)	41 (5.0)	42 (5.0)	39 (5.0)	47 (5.0)
70	64 (4.6)	70 (4.6)	66 (4.6)	66 (4.6)	55 (4.6)	63 (4.6)
90	86 (3.0)	87 (3.0)	88 (3.0)	89 (3.0)	74 (3.0)	82 (3.0)
	Coverage rate $\bar{C}$ (%)					
10	10 (0.3)	9 (0.2)	9 (0.3)	8 (0.5)	9 (0.6)	10 (0.6)
30	31 (0.8)	28 (0.6)	28 (0.8)	24 (1.4)	27 (1.7)	31 (1.7)
50	51 (1.2)	47 (1.0)	47 (1.2)	41 (2.0)	46 (2.5)	52 (2.4)
70	71 (1.2)	66 (1.2)	66 (1.4)	60 (2.3)	65 (2.8)	72 (2.4)
90	91 (0.7)	86 (1.3)	87 (1.0)	83 (1.7)	89 (1.8)	92 (1.2)

†See the text for details of the simulations. 20 000 BATWING outputs were generated for each data set, corresponding to 1.6×10^8 accept–reject steps, after 8×10^6 were discarded as burn-in. The effective sample size is defined at equation (6) and estimated with a cut-off at lag 200. $\bar{H}$ is the number of data sets for which the $100p\%$ equal-tailed posterior interval includes the true parameter value for that simulation. $\bar{C}$ is the mean over data sets of the proportion of outputs which lie in the $100p\%$ equal-tailed prior interval. The exact (binomial) SD is given in parentheses for $\bar{H}$; the SD for $\bar{C}$ in parentheses is estimated from the 100 data points.

Averaging over data sets generated via random draws from the prior we have

$$E(C_x) = \iint C_x \, P(x, \phi'') \, \mathrm{d}x \, \mathrm{d}\phi'' = \iint_l^u P(x, \phi) \, \mathrm{d}\phi \, \mathrm{d}x = \int_l^u P(\phi) \, \mathrm{d}\phi = p.$$

Since $E(C_x) = E(H_x) = p$, but $C_x \in [0, 1]$, it follows that $\mathrm{var}(C_x) \leqslant \mathrm{var}(H_x) = p(1 - p)$. This is verified by the lower part of Table 2, which shows $\bar{C}$, the average of C_x over the simulated data sets, together with its sample SD. Agreement between the achieved and nominal coverage as measured by $\bar{C}$ is reasonably good, but the discrepancy exceeds three SDs for μ and L. The largest discrepancy is just under 4.5 SDs.

Overall, our conclusion is that, even after making an informal allowance for multiple comparisons, there are some indications of imperfect mixing or convergence or validity, but these seem unlikely to have a large effect on inferences. For most parameters of interest there is good agreement between the achieved and nominal coverage. The coverage is likely to be better for actual data analyses, because of the possibility of longer runs and more careful monitoring of convergence than is feasible in a simulation study. However, because of its large size (the

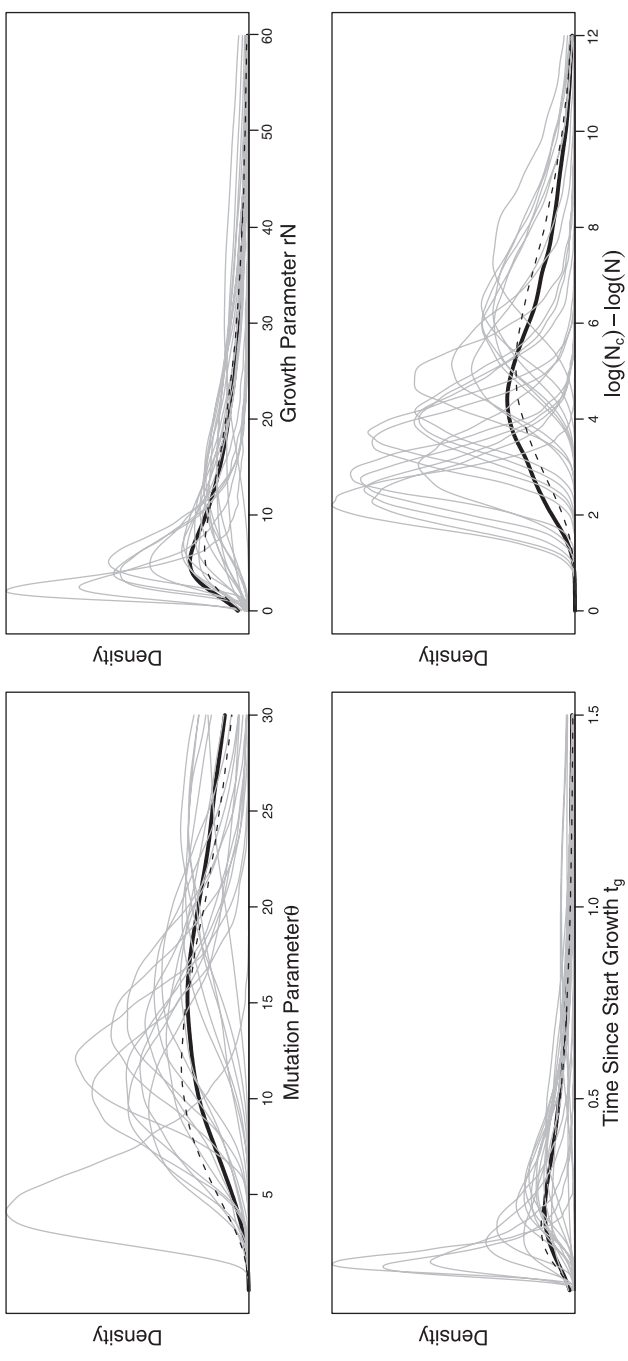

Fig. 5. Posterior density curves from the simulation study: in each panel, each of the 20 grey curves corresponds to a different simulated data set, chosen at random from the 100 data sets that were used in the simulation study; the broken curves indicate the prior density; the bold curves indicate the density obtained by combining 200 randomly chosen outputs for each data set (details of the BATWING runs are as for Table 2)

study occupied 10 computer processors for about 1 week) our simulation study was limited to a relatively small total sample size of 60 chromosomes.

Each panel of Fig. 5 shows posterior density curves estimated from the BATWING algorithm output for a random selection of 20 of the 100 simulated data sets. Perhaps the most noticeable feature of these curves is that, even in this ideal setting in which the modelling assumptions hold exactly, only weak inferences are possible for some demographic parameters. The reasons will be discussed further in Section 6. Fig. 5 also shows (bold curves) an average posterior density obtained by choosing a random selection of 200 outputs from each of the 100 data sets; in the limit of many data sets, this curve should coincide with the prior curve (broken curves). The agreement is generally good, but a discrepancy is apparent for μ and r.

5.2. Y-chromosome data

5.2.1. Modelling assumptions

5.2.1.1. Mutation at short tandem repeat loci. To formulate a prior for the STR mutation rate μ, Wilson and Balding (1998) drew on the Y-chromosome STR mutation study of Heyer *et al.* (1997), which recorded three mutations at tetranucleotide Y-STR loci from 1491 observed meioses. Adopting a gamma(1,1) (i.e. standard exponential) preprior, these data led Wilson and Balding (1998) to a gamma(4,1492) prior for μ, with mean $4/1492 \approx 2.7 \times 10^{-3}$. More data have since become available, and combining the results of Heyer *et al.* (1997) with those of Bianchi *et al.* (1998) and Kayser *et al.* (2000) we obtain a total of 17 mutations from 8169 meioses, suggesting a gamma(18,8170) prior for μ with mean $\approx 2.2 \times 10^{-3}$ (key quantiles of this and other prior distributions are shown in Table 3).

We have chosen to rely on the recent, directly observed data to formulate our prior distribution. Some researchers (e.g. Forster *et al.* (2000)) have questioned the use of contemporary pedigree data for historic mutation rates and point to indirect evidence for a lower mutation rate than is supported by our prior; see Siguroardottir *et al.* (2000) for a discussion of the issues. Pritchard *et al.* (1999) adopted a gamma(10,12500) prior, with mean $\approx 0.8 \times 10^{-3}$. Although this prior seems inconsistent with our choice, there is some overlap of the two prior density curves and the discrepancy is less than it first appears because Pritchard *et al.* (1999) allowed mutation steps of size greater than 1.

The STR loci of the C96 data set are all tetranucleotide repeats. Those of Ruiz Linares *et al.* (1996) are mostly dinucleotides, but include one of the tetranucleotides which was studied by Cooper *et al.* (1996) and all three of the mutation studies cited above. Some researchers have found dinucleotide mutation rates to be lower than tetranucleotides (Weber and Wong, 1993) whereas others have found the reverse relationship. Kayser *et al.* (2000) found no significant difference, although they have little dinucleotide data. Here, we adopt the simplest plausible model and assume a common mutation rate for dinucleotide and tetranucleotide STRs.

5.2.1.2. Y-chromosome Alu polymorphism locus. To allow at least partly for ascertainment bias, we condition on the existence of the YAP insert and assume *a priori* that it is equally likely to have occurred at any point on the genealogical tree. This assumption reflects the fact that the YAP locus was chosen to be typed because it was known in advance to be polymorphic, but it does not reflect the additional prior information that YAP is polymorphic in many human populations, and so cannot have arisen from a very recent mutation event. Although our prior could be modified to give more support to older times for the YAP mutation event, there is no obvious candidate for a specific assumption, and simulations indicate that more detailed modelling of the ascertainment scenario has no perceptible effect on posterior inferences (the

Table 3. Results from BATWING analyses of the C96 human Y-chromosome data set (Cooper *et al.*, 1996)†

Parameter	*Quantiles (5%, 50%, 95%)*					
	Prior			*Posterior*		
(a) Standard coalescent						
Effective population size N	2.0×10^3	4.7×10^3	9.2×10^3	2.5×10^3	4.0×10^3	6.4×10^3
Mutation rate μ per 10^3 generations	1.4	2.2	3.1	1.3	2.0	3.1
θ $(= 2N\mu)$	7.8	20	44	13	16	20
TMRCA (years)	63×10^3	200×10^3	600×10^3	21×10^3	43×10^3	120×10^3
(b) Coalescent with growth						
Ancestral population size N	820	2.7×10^3	6.3×10^3	170	770	2.0×10^3
Current effective population size N_c	15×10^3	280×10^3	28×10^6	12×10^3	27×10^3	84×10^3
Growth rate r (% per generation)	0.089	0.42	1.2	0.59	1.2	2.3
Mutation rate μ per 10^3 generations	1.4	2.2	3.1	0.90	1.5	2.4
Time since start growth, GNt_g (years)	7.2×10^3	28×10^3	150×10^3	4.3×10^3	7.4×10^3	14×10^3
TMRCA (years)	48×10^3	150×10^3	460×10^3	14×10^3	25×10^3	59×10^3
(c) Coalescent with splitting						
Effective population size N	2.0×10^3	4.7×10^3	9.2×10^3	2.8×10^3	4.1×10^3	7.1×10^3
Mutation rate μ per 10^3 generations	1.4	2.2	3.1	1.4	2.1	3.1
θ	7.8	20	44	15	19	23
Time since most recent split, GNt_b (years)	5.2×10^3	76×10^3	400×10^3	400	1.2×10^3	3.1×10^3
Time since 1st split, GNt_a (years)	33×10^3	180×10^3	670×10^3	5.1×10^3	10×10^3	21×10^3
TMRCA (years)	100×10^3	330×10^3	1.0×10^6	24×10^3	50×10^3	140×10^3
(d) Coalescent with splitting and growth						
Ancestral population size N	820	2.7×10^3	6.3×10^3	260	740	1.9×10^3
Current effective population size N_c	15×10^3	280×10^3	28×10^6	19×10^3	51×10^3	210×10^3
Growth rate r (% per generation)	0.089	0.42	1.2	0.58	1.2	2.1
Mutation rate μ per 10^3 generations	1.4	2.2	3.1	0.93	1.6	2.5
Time since start growth, GNt_g (years)	7.2×10^3	28×10^3	150×10^3	5.5×10^3	9.2×10^3	16×10^3
Time since most recent split, GNt_b (years)	2.5×10^3	42×10^3	260×10^3	1.8×10^3	4.1×10^3	7.9×10^3
Time since first split, GNt_a (years)	16×10^3	100×10^3	440×10^3	6.0×10^3	9.6×10^3	17×10^3
TMRCA (years)	60×10^3	200×10^3	660×10^3	16×10^3	29×10^3	64×10^3

†The prior distributions are as follows: $\mu \sim$ gamma(18,8170); $t_a \sim$ gamma(2,1); $t_b|t_a \sim$ uniform(0,t_a); for models (a) and (c) $N \sim$ gamma(5,10^{-3}); for models (b) and (d) $N \sim$ gamma(3,10^{-3}); $\log(N_c/N) \sim$ gamma(5,1); $r \sim$ gamma(2,400). The generation time $G = 25$ years. All estimates are based on 20000 BATWING outputs, corresponding to 1.6×10^8 accept–reject steps, after 8×10^6 were discarded as burn-in.

results are not shown). This is because the YAP data are informative about tree topology, but their weak effect on branch lengths is overwhelmed by information from the STR data.

5.2.1.3. Population size and growth parameters. Wilson and Balding (1998) investigated two prior distributions for N: lognormal(9,1) and gamma(5,10^{-3}). They found that the resulting posteriors of interest were very similar. For the models without population growth, we retain the gamma(5,10^{-3}) prior.

Estimates of the growth rates of human (census) population sizes over the past few thousand years can be obtained from the data reported by Cavalli-Sforza *et al.* (1994). From their data we estimate an average growth rate over the past few thousand years of about 1.2% per generation for the worldwide population, and 1.6% for Europe. However, *effective* population sizes can

be very different from census sizes, because of, for example, geographical stratification at various levels, age structure and mating behaviour. Plausibly, lower growth rates are appropriate for effective population sizes, and we adopt a gamma(2,400) prior for r, centred at 0.5% but supporting values above 1%.

Under constant population size models, the effective male population size over recent human evolution is often estimated to be of the order of 5000. It thus seems likely that N, the effective ancestral population size (before growth) was somewhat smaller and we choose a gamma$(3,10^{-3})$ prior. To specify a prior for the current effective population size, N_c, we assign the gamma(5,1) distribution to $\log(N_c/N)$. The prior for the time t_g at which growth started is implied by the priors described above, together with the relationship

$$\log(N_c/N) = rt_gN,$$

and the assumption that $\log(N_c/N)$, N and r are mutually independent.

5.2.1.4. Population subdivision parameters. We employ a gamma(2,1) prior for the time t_a at which the earliest split occurs. Subsequent splits occur at times which are jointly uniform, given t_a. The prior distribution of the subpopulation sizes, expressed as proportions of the total size, is Dirichlet$(2, 2, \ldots, 2)$. At each coalescence event in the population supertree, all possible coalescences are equally likely.

5.2.1.5. Generation time. Wilson and Balding (1998) assumed a value of 20 years for G, the male generation time. This may be plausible for females but is likely to be too low for males (see for example Tremblay and Vézina (2000)). Following Thomson *et al.* (2000), we adopt here $G = 25$ years. To facilitate a comparison with other studies, we do not model our uncertainty about the value of G. The data convey no information about G, so the posterior uncertainty is the same as the prior uncertainty.

5.2.1.6. Time since most recent ancestor. The prior distribution for TMRCA cannot be independently assigned. It depends (weakly) on the sample sizes but more importantly is a function of the demographic model and prior distributions for population sizes, growth rates and splitting times. Since the splitting time parameters arise in some demographic models and not others, it seems impossible in practice to specify essentially the same prior for TMRCA over our various models. In particular, allowing population growth leads to shorter values for TMRCA, whereas population splitting tends to increase these values, and only unrealistic priors for other parameters could fully compensate for these effects. However, our choices imply priors for TMRCA which overlap substantially, so that prior medians differ by a factor of less than 2 over the four demographic models, and the interval from 120 000 years to 450 000 years lies within the equal-tailed 90% interval for all four prior distributions for both sets of sample sizes (Tables 3 and 4).

5.2.2. Results: C96 data set

Wilson and Balding (1998) analysed two subsets of the C96 STR data. Here we analyse the full data set of 212 Y-chromosomes, described in Section 2.1, under all four demographic models introduced in Section 3.1. Key quantiles of the marginal posterior distributions under each of the four models are given in Table 3. In discussing these results we shall regard the posterior median of each parameter of interest as a point estimate.

Table 4. Results from BATWING analyses of the R96 data set of Ruiz Linares *et al.* (1996)†

Parameter	*Quantiles (5%, 50%, 95%)*					
	Prior			*Posterior*		
(a) Standard coalescent						
Effective population size N	2.0×10^3	4.7×10^3	9.2×10^3	2.4×10^3	3.8×10^3	6.3×10^3
Mutation rate μ per 10^3generations	1.4	2.2	3.1	1.3	2.0	3.1
θ $(= 2N\mu)$	7.8	20	44	12	15	19
TMRCA (years)	64×10^3	200×10^3	600×10^3	27×10^3	50×10^3	110×10^3
(b) Coalescent with growth						
Ancestral population size N	820	2.7×10^3	6.3×10^3	190	700	1.9×10^3
Current effective population size N_c	15×10^3	280×10^3	28×10^3	4.6×10^3	8.2×10^3	16×10^3
Growth rate r (% per generation)	0.089	0.42	1.2	0.19	0.44	0.94
Mutation rate μ per 10^3 generations	1.4	2.2	3.1	1.2	1.9	3.0
Time since start growth, GNt_g (years)	7.2×10^3	28×10^3	150×10^3	6.8×10^3	14×10^3	28×10^3
TMRCA (years)	48×10^3	150×10^3	450×10^3	17×10^3	31×10^3	63×10^3
(c) Coalescent with splitting						
Effective population size N	2.0×10^3	4.7×10^3	9.2×10^3	2.5×10^3	4.2×10^3	6.7×10^3
Mutation rate μ per 10^3 generations	1.4	2.2	3.1	1.3	2.1	3.1
θ	3.7	12	33	14	17	22
Time since most recent split, GNt_b (years)	560	11×10^3	86×10^3	0	60	320
Time since 1st split, GNt_a (years)	33×10^3	190×10^3	670×10^3	3.4×10^3	5.8×10^3	10×10^3
TMRCA (years)	120×10^3	360×10^3	1.0×10^6	27×10^3	51×10^3	110×10^3
(d) Coalescent with splitting and growth						
Ancestral population size N	820	2.7×10^3	6.3×10^3	140	540	1.7×10^3
Current effective population size N_c	15×10^3	280×10^3	28×10^6	7.9×10^3	15×10^3	32×10^3
Growth rate r (% per generation)	0.089	0.42	1.2	0.029	0.60	1.2
Mutation rate μ per 10^3 generations	1.4	2.2	3.1	1.1	1.8	2.8
Time since start growth, GNt_g (years)	7.2×10^3	28×10^3	150×10^3	7.4×10^3	14×10^3	26×10^3
Time since most recent split, GNt_b (years)	290	6.1×10^3	53×10^3	10	170	810
Time since 1st split, GNt_a (years)	16×10^3	100×10^3	440×10^3	5.0×10^3	8.1×10^3	13×10^3
TMRCA (years)	62×10^3	210×10^3	670×10^3	17×10^3	29×10^3	59×10^3

†The notation, prior distributions and details of the BATWING runs are as for Table 3.

The striking feature of almost all human Y-chromosome data, including the present data set, is its lack of variability, and the consequences of this are evident in the results reported in Table 3. Under the standard coalescent model, the estimated value of N is about 4000, which may seem implausibly low but is consistent with the results of other studies (Jorde *et al.*, 2001). The estimated TMRCA is 43 000 years, which is well below the 5% quantile of the prior distribution. This result is similar to that obtained by Wilson and Balding (1998), Pritchard *et al.* (1999) and Thomson *et al.* (2000), but until recently it would have been regarded as too low since, for example, humans arrived in Australia before that time. One goal of the present analyses is to investigate to what extent this unexpected result might be due to inappropriate modelling assumptions.

Weakening the assumptions of the standard coalescent model to allow for population growth leads to even less plausible results: the estimated TMRCA falls further to only 25 000 years. The growth rate estimate is large (1.2%), but the growth started only recently (estimate 7400 years). Pritchard *et al.* (1999) estimated a lower growth rate (0.8%) and a longer time since growth

(18000 years); our values lie within their 95% intervals. The current effective population size is estimated at only 27000, which is several orders of magnitude less than the census population size, but in accord with the estimate of 28000 obtained by Thomson *et al.* (2000) by using sequence data analysed via GENETREE (see Section 6).

However, extending the standard coalescent model to allow for geographical structuring via the splitting model (without growth) increases the estimates of both N and TMRCA by about 10%. The first population split occurred recently (median about 10000 years ago), and this was almost certainly the Nigerian population splitting from the common ancestral population of East Anglians and Sardinians (99.9% of MCMC outputs; the prior assigns probability $\frac{1}{3}$ to each possible split). The more recent split is estimated to have occurred approximately 1000 years ago. Care must be taken in interpreting these times since the splitting model does not allow migration after a split. Estimates would therefore reflect the end of a gradual splitting process and could be misleading if there had been substantial migration following a splitting event.

Adding growth to the splitting model, the most recent split moves further into the past, but the TMRCA estimate is reduced substantially (29000 years). Since the initial split, the growth rate (which is assumed common to all populations) is very high (estimate 1.6% per generation), but the estimate of the current population size remains low (51000) because of the small size of the ancestral population (estimate 740) and the short period of growth (estimate 9200 years). The probability that it was the Nigerian population which split first from the common ancestral population is now 99.1%, with the remaining 0.9% probability being roughly equally divided between the Sardinian and East Anglian populations. The posterior mean effective size of the Nigerian population is 37% of the total effective population size, with the East Anglian and Sardinian population proportions being 31% and 32% respectively.

5.2.3. Results: R96 data set

From the 121 Y-chromosomes of the R96 data set (Ruiz Linares *et al.*, 1996), we analysed the 115 chromosomes with no missing data at the YAP and SNP sites. Although missing STR data are relatively easy to handle, data missing from UEP sites are more problematic, and removing just six chromosomes seemed preferable to an arbitrary assumption. Sample sizes for this subset (and the full data set) are shown in Table 5.

Although the data set is very different from that of Cooper *et al.* (1996), some aspects of the posterior distributions summarized in Table 4 are strikingly similar. For example, the estimates of N under each model are similar for the two data sets, even though the number and distribution of source populations differ greatly. A possible interpretation is that levels of migration among human populations have been such that the current geographic location is relatively unimportant in explaining genetic diversity. More striking still is the similarity across the two data sets of the four TMRCA estimates.

One difference between the two sets of results is that the growth rate estimates are lower for the R96, worldwide, data set than for the C96 data set which is dominated by Europe.

The estimates of relative (effective) sizes of the 13 populations of the R96 data set (given in Table 5) are surprisingly uniform, with posterior means ranging from about 7% to 9% of the total, and bearing no apparent relationship with current census population sizes. The three largest values correspond to the three African populations, in accord with a greater genetic diversity in Africa than in other continents, which has been reported by many previous researchers. The two Amazonian populations had the smallest values.

Table 6 gives some probabilities for clusterings of different populations under the splitting model, with and without growth. The most recent population split was, with probability 38%,

Table 5. Sample sizes for the R96 human Y-chromosome data set (Ruiz Linares *et al*., 1996)†

Population or region	*Code*	*Sample size*		*Mean population proportions (%)*	
		(a)	*(b)*	*Splitting only*	*Splitting and growth*
Lisongo	LI	4	4	8.0	8.2
CAR‡ pygmy	PC	12	12	8.2	8.2
Zaire pygmy	PZ	9	10	8.5	8.8
Africa total	AFR	25	26	24.7	25.1
Karitania	KA	11	11	7.3	7.0
Maya	MA	9	9	8.0	7.9
Surui	SU	8	8	7.0	6.9
Americas total	AME	28	28	22.2	21.8
Cambodia	CA	15	15	7.5	7.4
China	CH	9	13	7.5	7.5
Japan	JA	11	11	7.9	7.8
East Asia total	ASI	35	39	22.9	22.6
Australia	AU	2	2	7.4	7.5
Melanesia	ME	4	4	7.4	7.6
New Guinea	NG	6	6	7.6	7.6
Oceania total	OCE	12	12	22.4	22.7
Europe	EU	15	16	7.8	7.7
Grand total		115	121	100.0	100.0

†Numbers of chromosomes (a) used in our analyses, having no missing data at the two UEP sites, and (b) in the original data set. Also shown are the posterior mean effective population sizes, as percentages of the total, under the two models which allow population splitting (the prior means are 7.7% for each population).
‡CAR, Central African Republic.

between Cambodia and China. The deepest split is estimated to have been between the African and non-African populations, a finding which has also been frequently reported from previous data sets. Although this split is assigned a posterior probability of only about 16%, note that the 13 populations are not grouped into continents *a priori*, so the $2^{13} - 1$ possible groupings at the root split are initially equally likely. Thus the data provide very strong support for the deepest split being between the African and non-African populations, as well as for the two other continental groupings shown in Table 6.

Looking at groupings which arise at any node in the tree, the three African populations form the strongest cluster, occurring in over 90% of trees. The two Amazonian populations also cluster together frequently. The strongest cross-continent clusterings involve New Guinea with the three Asian populations and Europe with the three American populations. The latter clustering seems surprising but may be due either to a substantial gene flow into both America and Europe from north or central Asia during and/or since the last ice age or perhaps very recent gene flow direct from Europe into native American populations during the period of European colonization. (The data sets consist of aboriginal peoples as far as this can be verified, but there may nevertheless be some recent admixture.)

Table 6. Posterior probabilities for various population groupings from the R96 data set†

Population grouping	*Posterior probability (%)*	
	Splitting only	*Splitting and growth*
(a) At the most recent split		
CA, CH	38	38
LI, PZ	20	18
(b) At the root split		
AFR *versus* non-AFR	17	15
AFR + ASI + OCE *versus* AME + EU	14	15
AFR + AME + EU *versus* ASI + OCE	7	10
(c) At any split		
AFR	95	91
KA + SU	69	79
CA + CH	63	65
LI + PZ	57	51
AME	48	42
ASI + NG	33	47
AME + EU	31	45
CA + CH + NG	46	45

†See Table 5 for the population codes. For (c), each entry gives the proportion of BATWING outputs such that the tree has a node which is ancestral to the stated populations only. The table shows all groupings with support under at least one model of 10% or more for (a) and (b), and 40% or more for (c)

5.3. β-globin sequences

5.3.1. *Modelling assumptions*

Harding *et al.* (1997) inferred an ancestral haplotype for their sample by comparison with a chimpanzee sequence. Using the standard coalescent model together with the infinite sites assumption, they obtained a point estimate of 2.55 for the mutation parameter θ. Conditional on this estimate, and on an estimate of the mutation rate obtained from human–chimpanzee comparisons, they inferred a point estimate of 895000 years for TMRCA (95% confidence interval 380000–1410000 years).

Our reanalyses regard the tree as unrooted *a priori* and draw inferences about the root without using the chimpanzee sequence. Further, we incorporate uncertainty about μ, the per sequence and per generation mutation rate. We started with a gamma$(3,10^5)$ preprior for μ, based on a genome-wide substitution rate estimate of 10^{-8} per site. Harding *et al.* (1997) observed 31 sites monomorphic in humans but varying between humans and chimpanzees. Assuming that there have been about 2.5×10^5 generations since the human–chimpanzee split, this leads to a gamma$(34,6\times10^5)$ distribution for μ, which we adopted as a prior distribution for our analyses. For the effective population size N, we assume respectively a gamma$(5,2.5\times10^{-4})$ and a gamma$(3,2.5\times10^{-4})$ prior distribution under the standard coalescent and the coalescent-with-growth models (these are the priors used for the Y-chromosome analyses, scaled up by a factor of 4).

5.3.2. *Results*

Prior and posterior quantiles, under both the standard coalescent and the coalescent-with-

Table 7. Results from BATWING analyses of the H97 β-globin data set (Harding *et al.*, 1997)†

Parameter	*Quantiles (5%, 50%, 95%)*					
	Prior			*Posterior*		
(a) Standard coalescent						
N	7.9×10^3	19×10^3	37×10^3	12×10^3	21×10^3	33×10^3
μ (per 10^5 generations)	4.2	5.6	7.4	4.3	5.7	7.3
θ ($\equiv 2N\mu$)	0.84	2.1	4.3	1.4	2.4	3.7
TMRCA (years)	240×10^3	780×10^3	2.4×10^6	690×10^3	1.2×10^6	2.1×10^6
(b) Coalescent with growth						
Ancestral N	3.2×10^3	11×10^3	25×10^3	9.8×10^3	17×10^3	28×10^3
N_c	59×10^3	1.1×10^6	100×10^6	70×10^3	560×10^3	18×10^6
r (%)	0.088	0.42	1.2	0.23	0.66	1.5
μ	4.2	5.6	7.4	4.3	5.7	7.4
t_g (years)	7.3×10^3	28×10^3	150×10^3	4.8×10^3	14×10^3	36×10^3
TMRCA (years)	140×10^3	480×10^3	1.6×10^6	620×10^3	1.1×10^6	2.0×10^6

†The notation and details of the BATWING runs are as for Table 3. The prior distributions are as follows: $\mu \sim$ gamma(34,6×10^5); for model (a), $N \sim$ gamma(5,2.5×10^{-4}); for model (b), $N \sim$ gamma(3,2.5×10^{-4}); $\log(N_c/N) \sim$ gamma(5,1); $r \sim$ gamma(2,400). The generation time $G = 25$ years.

growth models, are shown in Table 7. Our modelling assumptions are similar to those of Harding *et al.* (1997), including the adoption of the infinite sites mutation model, and in many respects our results are similar. We use a generation time $G = 25$ years, for comparison with our other analyses, whereas Harding *et al.* (1997) used $G = 20$ years. After allowing for this difference, our posterior median estimates of TMRCA (1 100 000 and 1 200 000 years ago) are similar to the point estimate of Harding *et al.* (1997), but our 90% interval is wider than even their 95% interval, because we model uncertainty about N, μ and the root of the tree. In addition, the confidence interval of Harding *et al.* (1997) is symmetric about the estimate (± 2 SDs), whereas our posterior distribution is skew and our interval is shifted towards higher values.

Table 8 shows part of the posterior distribution for the root node sequence, based on the human data alone. The most likely root sequence is the one assumed by Harding *et al.* (1997), based on the chimpanzee comparisons, indicating that such outgroup comparisons are not required to estimate ancestral sequences, although these are of course helpful if they are available.

Table 8. Posterior probabilities of the MRCA state for the H97 data set

Sequence	*Posterior probability of being root (%) for the following models:*	
	Standard coalescent	*Coalescent with growth*
TTTCCTCTGGCAT	17.5	16.5
TTTCCTCCGGCAT	6.9	6.2
TTTTCTCTGGCAT	6.7	6.8
TTTCCTCTGGAAT	6.4	6.5
TCTCCTCTGGCAT	6.3	6.7
TCTTCTCCGGAAT	6.0	6.2
All others	50.2	51.1

Table 9. Posterior quantiles of the age of each mutation for the H97 data set

Site	*Age of mutation ($\times 10^3$ years)*			
	H97	*Posterior quantiles*		
		5%	*50%*	*95%*
532	18	0.69	11	84
2634	61	6.7	71	240
2554	63	26	100	280
1423	100	26	100	280
379	84	34	110	270
1358	310	130	300	650
2792	530	120	330	810
2945	450	140	350	840
1416	460	140	350	840
2008	390	340	800	1600
906	510	340	800	1600
508	620	340	800	1600
2636	730	340	800	1600

Table 9 shows posterior quantiles for the age of each mutation (which is assumed unique at each segregating site under the infinite sites model). Our results are in broad agreement with those of Harding *et al.* (1997) except that we report (identical) marginal posterior distributions for mutations which cannot be distinguished from the data, whereas they reported, in effect, order statistics.

We also performed our analyses under a finite sites mutation model, which does not require the assumption of at most one mutation at each site. For these data, weakening the infinite sites model did not lead to any substantial changes in inferences (the results are not shown).

5.4. Mitochondrial minisequences

5.4.1. Modelling assumptions

Each of the three racial groups in the FSS mitochondrial DNA data set (Section 2.3) were analysed using the F84 mutation model for the 10 SNP sites, the SMM for the STR locus and the standard coalescent model, with the same prior distributions in each case.

The prior distributions that were adopted for the analysis are shown in Table 10. Prior distributions which are symmetric on the logarithmic scale seem appropriate for the ratios π_A/π_T, π_C/π_T and π_G/π_T. We chose prior medians that were close to the sample proportions, averaged over the 10 SNP sites studied here, in the data compiled by Handt *et al.* (1998), consisting of several thousand human mitochondrial DNA sequences (an updated version is available at `http://www.hvrbase.de`). Because of shared ancestry, the prior variances should be much higher than would be justified by sampling error in the background data. We chose variances such that the prior density was at least half the modal value within a factor of 2 either side of the median.

The database of Handt *et al.* (1998) was also used to formulate a prior for the ratio of *transversions* to transitions at these 10 sites. A rough point estimate of this ratio is obtained by taking

Table 10. Prior distributions for the mutation and demographic parameters in the mitochondrial DNA analysis†

Parameter	*Distribution*	*Quantiles*		
		2.5%	*50%*	*97.5%*
π_A/π_T	lognormal(−2.4,0.6)	0.028	0.091	0.29
π_C/π_T	lognormal(−0.61,0.6)	0.17	0.54	1.76
π_G/π_T	lognormal(−0.56,0.6)	0.18	0.57	1.85
β/α	lognormal(−4.5,2)	0.00022	0.011	0.56
$\alpha/2.5$	lognormal(−11,2)	0.033×10^{-5}	1.7×10^{-5}	84×10^{-5}
μ_{SNP}		0.031×10^{-5}	1.6×10^{-5}	88×10^{-5}
μ_{STR}	lognormal(−7,2)	0.18×10^{-4}	9.1×10^{-4}	460×10^{-4}
N	lognormal(9,1)	1100	8100	58000

†The prior for μ_{SNP} is determined by equation (5) and the priors above it in the table; for this row only, the quantiles stated are simulation-based approximations.

the most frequently observed nucleotide at each site as the ancestral state and comparing the number of haplotypes that represent a transversional change to the number that represent a transitional change. For the FSS data, this ratio is 1:127, which is considerably lower than estimates for the control region as a whole (Meyer *et al.*, 1999), which may reflect ascertainment bias in the selection of these sites. The prior chosen for β/α has a median that is close to this point estimate but also supports the higher estimates.

The prior for μ_{SNP} is implied by the definition (5) and the priors for the five parameters α, β, π_A/π_T, π_C/π_T and π_G/π_T. The priors defined so far suggest the approximation $\mu_{SNP} \approx \alpha/2.5$. The prior for $\alpha/2.5$ was chosen to be sufficiently diffuse to cover both the (low) estimates for μ_{SNP} derived from phylogenetic studies and the (high) estimates derived from pedigree and coalescent studies (Siguroardottir *et al.*, 2000).

The prior for μ_{STR} was chosen to cover the range of estimates for mammalian dinucleotide STRs (Weber and Wong, 1993; Schug *et al.*, 1998). Similarly, the prior for N covers published point estimates for the effective, ancestral population size of human mitochondrial DNA (Sherry *et al.*, 1997; Bonneuil, 1998).

For each of the racial groups, an MCMC-based approximation to the match probability under our model was obtained treating as the crime scene haplotype

(a) the haplotype that is most common in that racial group,
(b) a haplotype that is unobserved in that group but is very similar to the commonest haplotype and
(c) a haplotype that is unobserved and is also very dissimilar to any observed haplotype

(the crime scene haplotype differs from each observed haplotype by a transversion substitution at each SNP site).

Recently evidence has been presented for recombination in mitochondrial DNA, but the consensus seems to remain that it is not subject to recombination (Eyre-Walker and Awadalla, 2001). If this proves to be wrong, the implications for our results may not be great since recombination would presumably be rare and the sites that are studied here are all from the mitochondrial DNA control region.

Table 11. Match probabilities for three mitochondrial DNA haplotypes in each of three FSS databases†

Population	*Haplotype*		*P(match) (%)*		*Average tree length*	*Average tree height*	*Branch length*
	Type	*Description*	*Naïve*	*MCMC*			
Caucasian, $n = 153$	ATTTG5CGTTT	Common	25	18	9.6	1.4	0.018
	ATTTG6CGTTT	Similar	0.65	1.2	9.6	1.4	0.035
	CAGGT6ACGAG	Dissimilar	0.65	0.79	13	3.2	0.027
Afro-Caribbean, $n = 104$	GTCCA4CGCTC	Common	9.6	5.0	9.4	1.6	0.016
	GTCCA5CGCTC	Similar	0.96	0.51	9.3	1.5	0.024
	CAGGT6ACGAG	Dissimilar	0.96	0.93	13	3.3	0.059
Asian, $n = 43$	GTCTG5CGTTT	Common	21	16	8.5	1.9	0.059
	GTCTG4CGTTT	Similar	2.3	1.7	8.3	1.8	0.091
	CAGGT6ACGAG	Dissimilar	2.3	2.2	13	4.2	0.064

†The haplotypes chosen were the most common in the database ('common'), a haplotype which is unobserved in the database but which differs only at the STR locus from the most common haplotype ('similar') and an unobserved haplotype which is dissimilar to any observed haplotype ('dissimilar'). The match probabilities were calculated using the following: the relative frequency among the $n + 1$ observed haplotypes ('naïve'); the coalescent model with the F84 model and SMM, approximated from 10^5 MCMC outputs. The average height and total length of the genealogical trees are shown in coalescent time units, as is the length of the branch from x, the unobserved terminal node, to z, the most recent ancestor shared with an observed sequence.

5.4.2. Results

The MCMC-based match probabilities are shown under 'MCMC' in Table 11, along with some average properties of the genealogical tree. As expected, when the crime scene haplotype is dissimilar to any common haplotype, the genealogical tree is much higher than for a 'similar' haplotype. The effect on the match probability is, however, less predictable: for Caucasians, the dissimilar haplotype has a lower match probability than the similar haplotype, whereas the ordering is reversed for the other two populations. Two factors may contribute to this: the larger sample size for Caucasians and the fact that the most common Caucasian haplotype has a higher relative frequency than do the most common Asian and Afro-Caribbean haplotypes.

The relative frequency of the crime scene profile among the $n+1$ observed sequences provides a natural approximation to the match probability. It would be the maximum likelihood estimate of the population proportion if observed sequences were treated as independent, ignoring the genealogical structure and ascertainment (the profile of interest is the last one sampled). The number of possible mitochondrial DNA sequences is vast, and many sequences which occur in the population will not be represented in the reference sample; hence the relative frequency of the mitochondrial DNA profile of s may tend to overstate the match probability. In typical forensic applications, such an overstatement would favour defendants and may be regarded as less serious than an error which tends to disadvantage defendants. It is therefore of interest to investigate situations in which the relative frequency understates the match probability; if these are sufficiently rare then relative frequency approximation may suffice in place of a more carefully calculated match probability. Table 11 suggests that the naïve match probability will usually, but not always, be conservative.

The results presented here are based on the standard coalescent model. We have repeated the analyses using the coalescent-with-growth model and found no appreciable differences in the match probabilities (the results are not shown).

6. Discussion

Changes in population size and structure correspond to rescaling time in the coalescent tree. For example, relative to the standard coalescent model, the times between coalescence events are greater when the population size is large, and vice versa. Therefore the pattern of coalescence events provides indirect evidence about past demographic parameters. However, coalescences are not observed, but are inferred from the data. Moreover, these inferences are usually imprecise: if the mutation rate is high then all trace of the earliest coalescences may be eliminated by subsequent mutations, whereas if the mutation rate is low there will be insufficient mutations to 'document' many of the coalescences.

More generally, genealogical models such as coalescent models describe a complex stochastic process evolving through time, whereas most data sets, including those examined in the present paper, provide information at only a single time point. For some demographic parameters, more precise inferences can be made by using multiple unlinked loci, in which case every locus gives independent information (Beaumont, 1999). For genealogical parameters such as TMRCA, and for all inferences based on Y-chromosome and mitochondrial DNA data (which have their own unique male- and female-mediated demographies), this option is not available and only imprecise inferences can reasonably be expected on the basis of genetic information alone, even by using the most powerful methodology. However, genetic data can profitably be combined with information from other sources, such as palaeontology, archaeology and historical records.

The imprecision of inferences is reflected in the posterior distributions shown in Fig. 5, some of which are very diffuse even though the modelling assumptions hold exactly. Despite the poor prospects for precise answers, questions of human population history are of such intrinsic interest, and potential usefulness in understanding other aspects of human genetics, that they seem worth pursuing.

Although we cannot accurately estimate growth rates or splitting times from the data that were considered here, some inferences can be made with reasonable confidence. In particular, the low estimates of the TMRCA for Y-chromosome data seem reasonably robust to modelling assumptions. Our results are further supported by Thomson *et al.* (2000), who analysed different Y-chromosome data and employed different modelling assumptions and a different method of analysis, as well as by Pritchard *et al.* (1999), who analysed a larger data set of 445 chromosomes, typed at eight STRs, and used similar modelling assumptions, but employed a rejection method for approximating the posterior distribution, based on a vector of non-sufficient summary statistics.

In contrast with the results of the Y-chromosome analyses, the TMRCA estimates for the β-globin data are high relative to the prior distribution. In fact, the β-globin TMRCA is estimated to be very roughly 30-fold higher than the Y-TMRCA, whereas the simplest models would predict a fourfold difference. This may be due to homogenizing selection acting on the Y-chromosome (an advantageous Y-haplotype may have 'swept' through the population relatively recently), whereas balancing selection may be plausible for the β-globin locus, which would tend to increase TMRCA compared with expectations under a neutral model. However, there is substantial between-locus variability in TMRCA even under a neutral model, and so selection may not be needed to explain the observed disparity between Y-chromosome and β-globin results.

Alternative software is available, providing analyses that are similar in some respects to those performed by BATWING. Kuhner *et al.* (1995, 1998) and Beerli and Felsenstein (1999, 2001) have developed LAMARC, a suite of MCMC algorithms for the estimation of parameters such

as the exponential growth rate and migration rates, available at

```
http://evolution.genetics.washington.edu/lamarc.html
```

These algorithms do not exploit auxiliary variables at the internal nodes of the tree. They apply the MCMC algorithm only to the genealogical tree: the demographic and mutation parameters are fixed in each run, and a likelihood surface is inferred by using importance sampling re-weighting of a number of runs with different driving values. Stephens and Donnelly (2000) have pointed out problems with this approach: the variance of the importance weights can become large for points on the likelihood surface that are not close to the driving values.

GENETREE is described in Bahlo and Griffiths (2000) and is available at

```
http://www.stats.ox.ac.uk/~griff/software.html
```

It approximates likelihood surfaces for the mutation rate under the infinite sites mutation model, an exponential growth rate and a matrix of migration parameters under the island model. The methodology underlying GENETREE can be viewed as a version of importance sampling; the run times are often very long and Stephens and Donnelly (2000) have discussed methods for choosing more efficient importance sampling weights.

For forensic mitochondrial DNA match probabilities, we have found that the current practice of reporting sample relative frequencies seems to be conservative in most but not all cases, provided that the crime scene profile is included in the calculation (because the crime scene sequence is the last sampled, this implies that a zero frequency can never occur). Tully *et al.* (2001) discussed the more conservative approach of using the relative frequency after the crime scene profile has been added twice to the database, as well as the upper 95% confidence limit, which is even more conservative. For Y-chromosome match probabilities, Roewer *et al.* (2000) employed a posterior mean with respect to a prior beta distribution, with parameters determined by the observed haplotype diversity, which leads to less conservative values. None of the methods, including our own, includes modelling regional variation on a scale that is finer than that for which databases are available, and this may be important for both mitochondrial DNA and Y-chromosome match probabilities.

We believe that the methodology presented here represents an important advance towards the goal of fully likelihood-based methods for analysing DNA sequence data: we can obtain simultaneous inferences about a wider range of demographic, evolutionary and genealogical parameters, together with more realistic assessments of uncertainty, and for a wider range of DNA data types, than has hitherto been feasible. The present work, together with other recent developments, brings closer the goal of quantitative model criticism, and model comparisons, for detailed statistical models of the genetic history of humans and other species. Under our most complex model, we can currently analyse in reasonable time (say, a few days on a modest desk top workstation) sample sizes of over 200 chromosomes with haplotypes of up to about 10 STR markers (additional UEPs typically reduce the computation time and so are effectively unlimited). Further developments of algorithms and faster computers will continue to increase the feasible sample sizes.

Acknowledgements

We thank Andres Ruiz Linares, Rosalind Harding and Gillian Tully and Ian Evett of the UK FSS, for providing data and useful advice. Many thanks are due to the various users of BATWING for feed-back on the program, in particular Oliver Pybus, for helping with the

Macintosh version, and Noah Rosenberg. This work was supported by the UK Engineering and Physical Sciences Research Council under grant GR/K72599.

References

Adcock, G. J., Dennis, E. S., Easteal, S., Huttley, G. A., Jermlin, L. S., Peacock, W. J. and Thorne, A. (2001) Mitochondrial DNA sequences in ancient Australians: implications for modern human origins. *Proc. Natn. Acad. Sci. USA*, **98**, 537–542.

Aldous, D. J. (2000) Mixing time for a Markov chain on cladograms. *Combin. Probab. Comput.*, **9**, 191–204.

Allen, M., Engstrom, A. S., Meyer, S., Handt, O., Saldeen, T., von Haeseler, A., Paabo, S. and Gyllensten, V. (1998) Mitochondrial DNA sequencing of shed hairs and saliva on robbery caps: sensitivity and matching probabilities. *J. Forens. Sci.*, **43**, 453–464.

Anderson, E. C., Williamson, E. G. and Thompson, E. A. (2000) Monte Carlo evaluation of the likelihood for N_e from temporally spaced samples. *Genetics*, **156**, 2109–2118.

Bahlo, M. and Griffiths, R. C. (2000) Inference from gene trees in a subdivided population. *Theoret. Popln Biol.*, **57**, 79–95.

Balding, D. J., Bishop, M. and Cannings, C. (eds) (2001) *Handbook of Statistical Genetics*. Chichester: Wiley.

Balding, D. J. and Donnelly, P. (1995) Inferring identity from DNA profile evidence. *Proc. Natn. Acad. Sci. USA*, **92**, 11741–11745.

Bataille, M., Crainic, K., Leterreux, M., Durigon, M. and deMazancourt, P. (1999) Multiplex amplification of mitochondrial DNA for human and species identification in forensic evaluation. *Forens. Sci. Int.*, **99**, 165–170.

Beaumont, M. (1999) Detecting population expansion and decline using microsatellites. *Genetics*, **153**, 2013–2029.

Beaumont, M. (2001) Conservation genetics. In *Handbook of Statistical Genetics* (eds D. J. Balding, C. Cannings and M. Bishop), ch. 27. Chichester: Wiley.

Beerli, P. and Felsenstein, J. (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, **152**, 763–773.

Beerli, P. and Felsenstein, J. (2001) Maximum-likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proc. Natn. Acad. Sci. USA*, **98**, 4563–4568.

Bianchi, N. O., Catanesi, C. I., Bailliet, G., Martinez-Marignac, V. L., Bravi, C. M., Vidal-Rioja, L. B., Herrera, R. J. and Lopez-Camelo, J. S. (1998) Characterization of ancestral and derived Y-chromosome haplotypes of new world native populations. *Am. J. Hum. Genet.*, **63**, 1862–1871.

Bonneuil, N. (1998) Population paths implied by the mean number of pairwise nucleotide differences among mitochondrial DNA sequences. *Ann. Hum. Genet.*, **62**, 61–73.

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J. and Rolf, B. (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am. J. Hum. Genet.*, **62**, 1408–1415.

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

Cannings, C. (1974) The latent roots of certain Markov chains arising in genetics: I, Haploid models. *Adv. Appl. Probab.*, **6**, 260–290.

Cavalli-Sforza, L. L. and Cavalli-Sforza, F. (1995) *The Great Human Diasporas: the History of Diversity and Evolution*. Reading: Addison-Wesley.

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

Chikhi, L., Bruford, M. W. and Beaumont, M. A. (2001) Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics*, **158**, 1347–1362.

Clayton, D. (2000) Linkage disequilibrium mapping of disease susceptibility genes in human populations. *Int. Statist. Rev.*, **68**, 23–43.

Cooper, G., Amos, W., Hoffman, D. and Rubinsztein, D. C. (1996) Network analysis of human Y microsatellite haplotypes. *Hum. Molec. Genet.*, **5**, 1759–1766.

Cooper, G., Burroughs, N. J., Rand, D. A., Rubinsztein, D. C. and Amos, W. (1999) Markov Chain Monte Carlo analysis of human Y-chromosome microsatellites provides evidence of biased mutation. *Proc. Natn. Acad. Sci. USA*, **96**, 11916–11921.

Donnelly, P., Nordborg, M. and Joyce, P. (2001) Likelihoods and simulation methods for a class of nonneutral population genetics models. *Genetics*, **159**, 853–867.

Donnelly, P. and Tavaré, S. (1995) Coalescents and genealogical structure under neutrality. *A. Rev. Genet.*, **29**, 410–421.

Eyre-Walker, A. and Awadalla, P. (2001) Does human mtDNA recombine? *J. Molec. Evoln*, **53**, 430–435.

Fearnhead, P. and Donnelly, P. (2002) Approximate likelihood methods for estimating local recombination rates. *J. R. Statist. Soc.* B, **64**, 657–680.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Molec. Evoln*, **17**, 368–376.

Felsenstein, J. and Churchill, G. A. (1996) A hidden Markov model approach to variation among sites in rate of evolution. *Molec. Biol. Evoln*, **13**, 93–104.
Forster, P., Röhl, A., Lünnemann, P., Brinkmann, C., Zerjal, T., Tyler-Smith, C. and Brinkmann, B. (2000) A short Tandem Repeat-based phylogeny for the human Y chromosome. *Am. J. Hum. Genet.*, **67**, 182–196.
Fu, Y.-X. and Li, W.-H. (1999) Coalescing into the 21st century: an overview and prospects of coalescent theory. *Theoret. Popln Biol.*, **56**, 1–10.
Fullerton, S. M., Harding, R. M., Boyce, A. J. and Clegg, J. B. (1994) Molecular and population genetic analysis of allelic sequence diversity at the human β-globin locus. *Proc. Natn. Acad. Sci. USA*, **91**, 1805–1809.
Goldstein, D. B. (2001) Islands of linkage disequilibrium. *Nat. Genet.*, **29**, 109–111.
Goldstein, D. B. and Schlötterer, C. (eds) (1999) *Microsatellites: Evolution and Applications*. Oxford: Oxford University Press.
Griffiths, R. C. and Marjoram, P. (1997) An ancestral recombination graph. *IMA J. Math. Applic.*, **87**, 257–270.
Griffiths, R. C. and Tavaré, S. (1994) Simulating probability-distributions in the coalescent. *Theoret. Popln Biol.*, **46**, 131–159.
Hammer, M. F. (1994) A recent insertion of an *Alu* element on the Y chromosome is a useful marker for human population studies. *Molec. Biol. Evoln*, **11**, 749–761.
Handt, O., Meyer, S. and von Haeseler, A. (1998) Compilation of human mtDNA control region sequences. *Nucleic Acids Res.*, **26**, 126–129.
Harding, R. M., Fullerton, S. M., Griffiths, R. C. and Clegg, J. B. (1997) A gene tree for β-globin sequences from Melanesia. *J. Molec. Evoln*, **44**, suppl. 1, S133–S138.
Hartl, D. L. and Clark, A. G. (1997) *Principles of Population Genetics*, 3rd edn. Sunderland: Sinauer.
Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Molec. Evoln*, **22**, 160–174.
Heyer, E., Puymirat, J., Dieltjes, P., Bakker, E. and de Knijff, P. (1997) Estimating Y-chromosome specific microsatellite mutation frequencies using deep rooting phylogenies. *Hum. Molec. Genet.*, **6**, 799–803.
Hudson, R. R. (1991) Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology* (eds D. J. Futuyama and J. Antonovics). Oxford: Oxford University Press.
Jorde, L. B., Watkins, W. S. and Bamshad, M. J. (2001) Population genomics: a bridge from evolutionary history to genetic medicine. *Hum. Molec. Genet.*, **10**, 2199–2207.
Jukes, T. H. and Cantor, C. R. (1969) Evolution of protein molecules. In *Mammalian Protein Metabolism* (ed. H. N. Munro), pp. 21–132. New York: Academic Press.
Kayser, M., Roewer, L., Hedman, M., Henke, L., Henke, J., Brauer, S., Kruger, C., Krawczak, M., Nagy, M., Dobosz, T., Szibor, R., de Kniff, P., Stoneking, M. and Sajantila, A. (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y-chromosome, as revealed by direct observation in father/son pairs. *Am. J. Hum. Genet.*, **66**, 1580–1588.
Kingman, J. F. C. (1982) The coalescent. *Stoch. Process. Applic.*, **13**, 235–248.
Krings, M., Stone, A., Schmitz, R. W., Krainitzki, H., Stoneking, M. and Paabo, S. (1997) Neandertal DNA sequences and the origin of modern humans. *Cell*, **90**, 19–30.
Kuhner, M. K., Yamato, J. and Felsenstein, J. (1995) Estimating effective population size from sequence data using Metropolis-Hastings sampling. *Genetics*, **140**, 1421–1430.
Kuhner, M. K., Yamato, J. and Felsenstein, J. (1998) Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, **149**, 429–434.
Kuhner, M. K., Yamato, J. and Felsenstein, J. (2000) Maximum likelihood estimation of recombination rates from population data. *Genetics*, **156**, 1393–1401.
Liu, J. S. (2002) *Monte Carlo Strategies in Scientific Computing*. New York: Springer.
Marjoram, P. and Donnelly, P. (1994) Pairwise comparisons of Mitochondrial DNA sequences in subdivided populations and implications for early human evolution. *Genetics*, **136**, 673–683.
Markovtsova, L., Marjoram, P. and Tavaré, S. (2000a) The age of a unique event polymorphism. *Genetics*, **156**, 401–409.
Markovtsova, L., Marjoram, P. and Tavaré, S. (2000b) The effects of rate variation on ancestral inference in the coalescent. *Genetics*, **156**, 1427–1436.
Mau, B., Newton, M. A. and Larget, B. (1999) Bayesian phylogenetic inference via Markov Chain Monte Carlo methods. *Biometrics*, **55**, 1–12.
Meyer, S., Weiss, G. and von Haeseler, A. (1999) Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics*, **152**, 1103–1110.
Nielsen, R. (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, **154**, 931–942.
Nielsen, R. and Wakeley, J. (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.
Nordborg, M. (2001) Coalescent theory. In *Handbook of Statistical Genetics* (eds D. J. Balding, C. Cannings and M. Bishop), ch. 7. Chichester: Wiley.
Ovchinnikov, I. V., Gotherstrom, A., Romanova, G. P., Kharitonov, V. M., Liden, K. and Goodwin, W. (2000) Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature*, **404**, 490–493.

Pfeiffer, H., Steighner, R., Fisher, R., Mornstad, H., Yoon, C. L. and Holland, M. M. (1998) Mitochondrial DNA extraction and typing from isolated dentin—experimental evaluation in a Korean population. *Int. J. Leg. Med.*, **111**, 309–313.
Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. and Feldman, M. W. (1999) Population growth of human Y chromosomes: a study of Y-chromosome microsatellites. *Molec. Biol. Evoln*, **16**, 1791–1798.
Relethford, J. (1998) Genetics of modern human origins and diversity. *A. Rev. Anthropol.*, **27**, 1–23.
Roewer, L., Kayser, M., de Knijff, P., Anslinger, K. and Betz, A. (2000) A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forens. Sci. Int.*, **114**, 31–43.
Rubin, D. B. and Schenker, N. (1986) Efficiently simulating the coverage properties of interval estimates. *Appl. Statist.*, **35**, 159–167.
Ruiz Linares, A., Nayar, K., Goldstein, D. B., Hebert, J. M., Seielstad, M. T., Underhill, P. A., Lin, A. A., Feldman, M. W. and Cavalli-Sforza, L. L. (1996) Geographic clustering of human Y-chromosome haplotypes. *Ann. Hum. Genet.*, **60**, 401–408.
Schug, M. D., Wetterstrand, K. A., Gaudette, M. S., Lim, R. H., Hutter, C. M. and Aquadro, C. F. (1998) The distribution and frequency of microsatellite loci on Drosophila Melanogaster. *Molec. Evoln*, **7**, 57–70.
Sherry, S. T., Harpending, H. C., Batzer, M. A. and Stoneking, M. (1997) *Alu* evolution in human populations: using the coalescent to estimate effective population size. *Genetics*, **147**, 1977–1982.
Siguroardottir, S., Helgason, A., Gulcher, J. R., Stefansson, K. and Donnelly, P. (2000) The mutation rate in the human mtDNA control region. *Am. J. Hum. Genet.*, **66**, 1599–1609.
Stephens, M. (2001) Inference under the coalescent. In *Handbook of Statistical Genetics* (eds D. J. Balding, C. Cannings and M. Bishop), ch. 8. Chichester: Wiley.
Stephens, M. and Donnelly, P. (2000) Inference in molecular population genetics (with discussion). *J. R. Statist. Soc.* B, **62**, 605–635.
Sykes, B. (1999) The molecular genetics of European ancestry. *Phil. Trans. R. Soc. Lond.* B, **354**, 131–139.
Thomson, R., Pritchard, J. K., Shen, P., Oefner, P. J. and Feldman, M. W. (2000) Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc. Natn. Acad. Sci. USA*, **97**, 7360–7365.
Tremblay, M. and Vézina, H. (2000) New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *Am. J. Hum. Genet.*, **66**, 651–658.
Tully, G., Bär, W., Brinkmann, B., Carracedo, A., Gill, P., Morling, N., Parson, W. and Schneider, P. (2001) Considerations by the European DNA Profiling (EDNAP) Group on the working practices, nomenclature and interpretation of mitochondrial DNA profiles. *Forens. Sci. Int.*, **124**, 83–91.
Tully, G., Sullivan, K. M., Nixon, P., Stones, R. E. and Gill, P. (1996) Rapid detection of mitochondrial sequence polymorphisms using multiplex solid-phase fluorescent minisequencing. *Genomics*, **34**, 107–113.
Weber, J. L. and Wong, C. (1993) Mutation of human short tandem repeats. *Hum. Molec. Genet.*, **2**, 1123–1128.
Weir, B. S. and Cockerham, C. C. (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.
Wilkinson-Herbots, H. M. (1998) Genealogy and subpopulation differentiation under various models of population structure. *J. Math. Biol.*, **37**, 535–585.
Wilson, I. J. and Balding, D. J. (1998) Genealogical inference from microsatellite data. *Genetics*, **150**, 499–510.
Wright, S. (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.

Discussion on the paper by Wilson, Weale and Balding

Ziheng Yang (*University College London*)

It is my great pleasure to propose the vote of thanks. The past two decades have seen a rapid accumulation of genetics data, and the focus of theoretical population genetics has shifted from forward mathematical modelling to inference from real data. The introduction of the coalescent model, 'a time machine which runs evolution backwards' (Edwards, 1970), and the development of computation-intensive statistical methods, such as Markov chain Monte Carlo (MCMC) and importance sampling, have made it possible to implement likelihood-based inference methods under biologically interesting models. In this paper, the authors describe their flexible MCMC algorithm for Bayes inference, which implements the standard coalescent model as well as models of deterministic population size change and population subdivision. The computer program handles all major types of genetics data, such as DNA sequences, short tandem repeats, single-nucleotide polymorphisms and unique event polymorphisms. It will no doubt become a powerful tool for population genetics analysis.

I would like to draw attention to some related work and to make two comments on the authors' algorithm. The population split model of the authors does not allow gene flow (migration) after the split, and it is quite similar to the model for estimating ancestral population sizes on a species phylogeny by using data from multiple loci. A maximum likelihood method was developed by Takahata *et al.* (1995) for two or three species under the infinite sites model, and Yang (2002) and Rannala and Yang (2003)

Pfeiffer, H., Steighner, R., Fisher, R., Mornstad, H., Yoon, C. L. and Holland, M. M. (1998) Mitochondrial DNA extraction and typing from isolated dentin—experimental evaluation in a Korean population. *Int. J. Leg. Med.*, **111**, 309–313.
Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. and Feldman, M. W. (1999) Population growth of human Y chromosomes: a study of Y-chromosome microsatellites. *Molec. Biol. Evoln*, **16**, 1791–1798.
Relethford, J. (1998) Genetics of modern human origins and diversity. *A. Rev. Anthropol.*, **27**, 1–23.
Roewer, L., Kayser, M., de Knijff, P., Anslinger, K. and Betz, A. (2000) A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forens. Sci. Int.*, **114**, 31–43.
Rubin, D. B. and Schenker, N. (1986) Efficiently simulating the coverage properties of interval estimates. *Appl. Statist.*, **35**, 159–167.
Ruiz Linares, A., Nayar, K., Goldstein, D. B., Hebert, J. M., Seielstad, M. T., Underhill, P. A., Lin, A. A., Feldman, M. W. and Cavalli-Sforza, L. L. (1996) Geographic clustering of human Y-chromosome haplotypes. *Ann. Hum. Genet.*, **60**, 401–408.
Schug, M. D., Wetterstrand, K. A., Gaudette, M. S., Lim, R. H., Hutter, C. M. and Aquadro, C. F. (1998) The distribution and frequency of microsatellite loci on Drosophila Melanogaster. *Molec. Evoln*, **7**, 57–70.
Sherry, S. T., Harpending, H. C., Batzer, M. A. and Stoneking, M. (1997) *Alu* evolution in human populations: using the coalescent to estimate effective population size. *Genetics*, **147**, 1977–1982.
Siguroardottir, S., Helgason, A., Gulcher, J. R., Stefansson, K. and Donnelly, P. (2000) The mutation rate in the human mtDNA control region. *Am. J. Hum. Genet.*, **66**, 1599–1609.
Stephens, M. (2001) Inference under the coalescent. In *Handbook of Statistical Genetics* (eds D. J. Balding, C. Cannings and M. Bishop), ch. 8. Chichester: Wiley.
Stephens, M. and Donnelly, P. (2000) Inference in molecular population genetics (with discussion). *J. R. Statist. Soc.* B, **62**, 605–635.
Sykes, B. (1999) The molecular genetics of European ancestry. *Phil. Trans. R. Soc. Lond.* B, **354**, 131–139.
Thomson, R., Pritchard, J. K., Shen, P., Oefner, P. J. and Feldman, M. W. (2000) Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc. Natn. Acad. Sci. USA*, **97**, 7360–7365.
Tremblay, M. and Vézina, H. (2000) New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *Am. J. Hum. Genet.*, **66**, 651–658.
Tully, G., Bär, W., Brinkmann, B., Carracedo, A., Gill, P., Morling, N., Parson, W. and Schneider, P. (2001) Considerations by the European DNA Profiling (EDNAP) Group on the working practices, nomenclature and interpretation of mitochondrial DNA profiles. *Forens. Sci. Int.*, **124**, 83–91.
Tully, G., Sullivan, K. M., Nixon, P., Stones, R. E. and Gill, P. (1996) Rapid detection of mitochondrial sequence polymorphisms using multiplex solid-phase fluorescent minisequencing. *Genomics*, **34**, 107–113.
Weber, J. L. and Wong, C. (1993) Mutation of human short tandem repeats. *Hum. Molec. Genet.*, **2**, 1123–1128.
Weir, B. S. and Cockerham, C. C. (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.
Wilkinson-Herbots, H. M. (1998) Genealogy and subpopulation differentiation under various models of population structure. *J. Math. Biol.*, **37**, 535–585.
Wilson, I. J. and Balding, D. J. (1998) Genealogical inference from microsatellite data. *Genetics*, **150**, 499–510.
Wright, S. (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.

Discussion on the paper by Wilson, Weale and Balding

Ziheng Yang (*University College London*)

It is my great pleasure to propose the vote of thanks. The past two decades have seen a rapid accumulation of genetics data, and the focus of theoretical population genetics has shifted from forward mathematical modelling to inference from real data. The introduction of the coalescent model, 'a time machine which runs evolution backwards' (Edwards, 1970), and the development of computation-intensive statistical methods, such as Markov chain Monte Carlo (MCMC) and importance sampling, have made it possible to implement likelihood-based inference methods under biologically interesting models. In this paper, the authors describe their flexible MCMC algorithm for Bayes inference, which implements the standard coalescent model as well as models of deterministic population size change and population subdivision. The computer program handles all major types of genetics data, such as DNA sequences, short tandem repeats, single-nucleotide polymorphisms and unique event polymorphisms. It will no doubt become a powerful tool for population genetics analysis.

I would like to draw attention to some related work and to make two comments on the authors' algorithm. The population split model of the authors does not allow gene flow (migration) after the split, and it is quite similar to the model for estimating ancestral population sizes on a species phylogeny by using data from multiple loci. A maximum likelihood method was developed by Takahata *et al.* (1995) for two or three species under the infinite sites model, and Yang (2002) and Rannala and Yang (2003)

implemented MCMC algorithms for Bayes inference for general species trees under finite sites models. Two observations that were made in those studies appear relevant here. First, simple methods based on summary statistics (such as the proportion of loci at which the gene tree does not match the species tree) may not use the information in the data efficiently and can be very misleading. Second, to estimate parameters in a complex model incorporating multiple factors, it seems necessary to combine data from multiple loci.

The first of my comments concerns the robustness of the posterior to the prior and to model assumptions. To some extent, one might argue that the authors' models are overparameterized. The simplest (standard coalescent) model has two parameters N and μ, whereas the likelihood depends on their product (θ) only. The model is not identifiable if uniform rather than gamma priors are used for N and μ. Although specifying priors for both parameters is a natural way of accounting for uncertainties in μ when we are estimating N, I wonder to what extent we are simply obtaining what we put in (both through the prior and through the model that is assumed). Overparameterization may produce slow convergence in the MCMC algorithm and strong correlation between parameters in the posterior, in which case the marginal credibility intervals are not an adequate summary of the joint posterior. The problem should be more serious under more sophisticated models of population growth and structure. I would like to see some comments on those issues. A way forward seems to be a combined analysis of data from multiple loci (nuclear, mitochondrial and Y-chromosome) to estimate the shared parameters such as the population growth rate and splitting times, while accommodating differences between loci (in mutation rate, population size, etc.).

My second comment concerns the proposal algorithm. We must make many quite arbitrary decisions, and it is often unclear which make an efficient algorithm. I am particularly interested in how the authors' data augmentation approach, which averages over ancestral states (at ancestral nodes in the genealogy) during the MCMC run, compares with the alternative approach of calculating that average directly by using the peeling or pruning algorithm (Felsenstein, 1981). The latter computation is linear with the sample size n even though the space of the ancestral states grows exponentially with n. A further saving is achieved if the proposal alters only parts of the genealogical tree, as duplicated computation in the unchanged parts is avoided. The posterior distribution of the ancestral states, if needed, can easily be recovered as well, at least for the root. Let the data be x, the ancestral states be y and the parameters in the model be θ. $p(y|x, \theta)$ is easily calculated from the pruning algorithm (Yang *et al.*, 1995), and

$$p(y \mid x) = \int p(y, \theta \mid x)\, \mathrm{d}\theta = \int f(y \mid x, \theta)\ f(\theta \mid x)\, \mathrm{d}\theta$$

can be calculated by averaging over the MCMC sample. I can see that for microsatellite data (short tandem repeats) the authors' use of ancestral states to generate proposals may increase the acceptance rate and the efficiency of the algorithm. However, this is less straightforward to implement for sequence data, and tying the proposal step (to change trees) with the mutation model or data type might complicate the algorithm. I would be delighted to hear any comments on the relative efficiency of those two strategies.

To conclude, the authors have produced a powerful and versatile program package that can accommodate different types of genetics data under several important population genetic models. I congratulate the authors for this achievement and have great pleasure in proposing the vote of thanks.

David Stephens (*Imperial College London*)

This is a very interesting paper that draws together aspects of many of the previously published methods of analysis and describes the implementation of the authors' own software in the analysis of some common types of DNA sequence data. The authors use extensions of the standard coalescent model and use a Markov chain Monte Carlo (MCMC) algorithm to analyse a variety of human-derived DNA data sets, thus extending the ground breaking paper of Wilson and Balding (1998). The modelling extensions in particular are very important, and the authors demonstrate that their algorithms can cope with the more complex models, albeit at some increased computational burden. The authors are to be congratulated on the major achievement of the development of a robust, accessible and efficient computational package.

I think that several important issues are raised by the paper.

The models

Models for the different data sets, in increasing order of complexity, the genealogical models used, are the standard coalescent (fixed population size, with population growth) or splitting–coalescent (with and without populations growth), the coalescent with population splitting and the coalescent with population

splitting and growth. Thus we have a rich and comprehensive treatment set of (nested) models, within which some key parameters are interpretable across all models. The prior specification for parameters in the models appears, however, to be quite problematic: for example, it is not clear that the prior specifications under the different growth models are sufficiently comparable to allow a straightforward posterior interpretation; contrast, for example, the prior and posterior for the TMRCA value for each model. Is it possible routinely to calibrate the priors for ease of comparison?

The algorithm

The algorithm that is used is an MCMC algorithm implemented through the BATWING package, using standard approaches to such MCMC problems. The authors clearly prefer the augmented likelihood approach, where the MCMC algorithm traverses the joint parameter–coalescent tree space, and it is clear that this offers a broader range of modelling possibilities. Do the authors have an opinion on the inferential and algorithmic performance of their method compared with non-MCMC methods?

Model selection and validation

The results that are presented for the various data sets raise the question of model selection and validation. For example, for Table 2, and given the results presented, what should the geneticist infer about TMRCA for the C96 data? We have the results for a range of models, but which set of results is most appropriate? Can population growth models be verified independently, with some form of time-stamped data?

Overall, for all the data sets, there is little or no discussion of model validity or comparison for the various models proposed; there is now the facility to fit sophisticated population genetics models to such sequence data but no real guide on how to compare the adequacy or otherwise of the utilized models *a posteriori*. When presented with, for example, the wealth of results in Tables 2, 4 or 6, what inference should the practitioner draw? The subjective Bayesian approach is, of course, completely coherent and to be recommended, but this does not remove the need for a thorough investigation of the effect of a range of prior specifications, or a *post hoc* model validation or assessment exercise. Although prior parameters are carefully chosen in each example, often on the basis of genuine prior opinion or historical data, the sensitivity of inference

(a) to the prior specification and
(b) any individual datum sequence

is not really discussed. In addition, the posterior behaviour as the sample size n changes may be of some interest, especially when n is small. Could the authors comment on, for example, the utility of bootstrap resampling—which is common in phylogenetics—or leave-one-out validation, population subsampling etc. to examine the stability of the posterior?

At the moment I am left with the feeling that I have no real idea about which of the models proposed (population structures and prior specifications) best represents the data. Much attention is given to understanding the convergence properties of the MCMC algorithm; I would regard the validity of the inference to be equally important. Current simulation-based Bayesian inference provides several different methods for assessing and comparing the fits of different models. First (approximations to) marginal likelihood quantities, the calculation of which for coalescent models are discussed and described in Stephens and Donnelly (2000) for example, can be used, and the model with the highest marginal likelihood preferred. Secondly, variable dimension MCMC methods can be used to compute posterior model probabilities. Neither method would impose a tremendous computational burden (with a slightly amended algorithm), but, I suspect, would detect any serious inconsistencies in prior or model specification.

The issue of identifiability—which parameters are inferable from the data—is not discussed at any length in the paper, but it is widely acknowledged that some parameters will always be estimated poorly (as they are only technically and not practically identifiable from the data alone). Apart from two parameters that are well known to be aliased (N and μ), are there any other parameters that display a similar strong dependence? This would be detected by inspection of joint posterior sample plots; none are included in the paper, but I assume that the authors have used such plots—if, for example, the TMRCA parameter is strongly correlated with other parameters in the posterior, reporting marginal results is questionable.

The simulation study is an attempt to verify consistency of the posterior, i.e. whether or not Bayesian posterior analysis regularly derives the correct result. The answer seems, generally, to be yes. What may be informative here would be to study performance for varying sample size; presumably (one hopes), in these simulations (taking, for example, $n = 30$ or $n = 120$ for the sample size), the principal reason that the posterior interval does not include the true value of the parameter is that the sample size is quite small. In the absence of technical results describing the asymptotic behaviour of the posterior it may be useful to

see what happens when the sample size is gradually increased beyond the practically realistic values that are selected in the paper.

In summary, I congratulate the authors on their significant contribution; the paper draws together several previously proposed methods of MCMC analysis for DNA sequence data and makes modelling and some algorithmic advances. Geneticists now have an array of models and computational algorithms with which to analyse their data. At this stage, there is little guidance on how to assess whether their inferences about the unobserved parameters of interest are credible in the light of the observed data, and conditional on all aspects of their model specification. Nevertheless, the work will be of considerable use and interest to geneticists and statisticians; I gladly second the vote of thanks to the authors.

The vote of thanks was passed by acclamation.

Kevin J. Dawson (*Rothamsted Research, Harpenden*)

The authors have extended the earlier approach of Wilson and Balding (1998) and Beaumont (1999) to allow for data sets where individuals have been sampled from separate populations and have incorporated the 'phylogeny' of these subpopulations, the 'supertree', as a parameter of the model. The assignment of individuals to contemporary subpopulations is also treated as a parameter of the model, about which we are uncertain. Each branch of the supertree is also associated with an *effective population size*. Migration between branches of the supertree (i.e. between ancestral subpopulations) is not allowed under the present model. I hope that this aspect of reality will be incorporated in the model in the near future.

The effect of selection on the genealogical process at loci which are tightly or loosely linked, or even unlinked, to the targets of selection (Barton, 1998) means that parameters of the genealogical process are no longer strictly determined by demography and should be treated as locus-specific parameters, or at least as parameters specific to tightly linked regions of the genome. It makes sense to combine information across marker loci on the Y-chromosome, as the authors have done, since these completely linked loci share the same gene genealogy. The assumption of complete linkage probably also applies to the mitochondrial genome. In contrast, the gene genealogies at unlinked or loosely linked autosomal (and X-chromosome) genes can be assumed to be statistically independent (unless the sample is drawn from a very small or otherwise closely inbred population). Here we should be much more cautious about assuming common values for effective population sizes across loci. The discrepancy between the inferences based on the Y-chromosome loci and the (autosomal) β-globin locus illustrate this point. Frequentist methods have been developed for identifying loci which have *outlying* genealogical histories (Bowcock *et al.*, 1991; Beaumont and Nichols, 1996; Vitalis *et al.*, 2001). It would be preferable to make these decisions within a Bayesian framework to make full use of the information provided by the data.

I presume that we should take the TMRCA of 29000 years BP for the Y-chromosome, based on the most general model, as being the more reliable estimate. The problem of reconciling this with the much earlier data for the colonization of Australia (between 40000 and 30000 years BP) is intriguing. The solution offered by the authors appears to be a selective sweep at the Y-chromosome, which could have extended worldwide, reaching Australia some time after 29000 years BP. This seems plausible. Have the authors considered how long it might have taken for such a selective sweep on the Y-chromosome to spread worldwide, or whether several geographically more restricted selective sweeps could have been responsible? How committed are they to such a recent TMRCA for the Y-chromosome?

Alexei Drummond (*University of Oxford*) **and Geoff Nicholls** (*University of Auckland*)

In elaborating a Bayesian Metropolis–Hastings Markov chain Monte Carlo (MCMC) framework for coalescent-based inference the authors provide an attractive alternative to both importance sampling (Griffiths and Tavare, 1994) and maximum likelihood MCMC methods (Kuhner *et al.*, 1995). Our comments arise from insight gained from our own published work on coalescent-based Bayesian MCMC kernels (Drummond *et al.*, 2002). The authors used data augmentation of sequences at internal nodes rather than the standard analytical peeling algorithm (Felsenstein, 1981). Data augmentation allows more complicated models of mutation and likelihood calculations are simplified. They studied small data sets, with only 13 variable sites in the H97 data set, leading to a small state space of ancestral sequences to sample. However, for larger more variable data sets, peeling will certainly be preferable, especially if ancestral sequences are nuisance parameters. Roughly speaking, MCMC sampling is slowed by diffuse distributions. When mutation rates are low, the distribution over ancestral sequences on a fixed tree is concentrated on a small set. At high mutation rates the MCMC algorithm must explore a relatively large set of ancestral sequences on each tree. In contrast the work in peeling is fixed. In our studies on temporally spaced leaf

data, we made comparable MCMC programs with and without peeling. Peeling had a clear advantage on large data sets. For sequences on a tree, it is straightforward to implement peeling so that local MCMC tree operations generate likelihood calculations that have $O\{\log(n)\}$ time complexity rather than $O(n)$ (where n is the number of leaves). Therefore, peeling is not as slow per update as might be expected.

A second consideration involves the separation of Θ into effective population size N_e and mutation rate μ. For contemporaneous sequences with no external calibrations of time, N_e and μ are confounded. This is true in the authors' work, and all information about N_e and μ beyond their product derives from the priors. Hence, their posterior of μ is almost identical to their prior (Tables 3, 4 and 7).

Finally, there is often doubt about what our state of knowledge actually is (so which prior we should use) and doubt about how to represent our knowledge mathematically (for high dimensional priors it is easy to write down a prior which, when sampled, produces typical realizations that are dramatically unrepresentative of our prior knowledge). Because of the conflation of N_e and μ, the primary achievement of this study is to give a method for converting prior knowledge about N_e and μ into knowledge about the timing of human origins, with an explicit quantification of uncertainty.

R. C. Griffiths (*Oxford University*)

The Melanesian data set that is considered by the authors is interesting in that it conforms to the infinitely many sites model of mutation with no recombination. The data are then equivalent to an essentially unique gene tree, constructed as a perfect phylogeny, whose vertices are labelled by mutations. Labelling of vertices is unique up to permutations of mutation labels along single edges. There is mathematical detail about the tree nature of this data set and ages of mutations in Griffiths and Tavaré (1999). Questions relating to the stochastic nature of the ancestral gene tree back in time can then be asked and answered by simulating trees back in time conditional on the topology of the gene tree by using computationally intensive methods with a combination of sequential importance sampling, Markov chain Monte Carlo or Bayesian methods. Fig. 6 shows an average gene tree from the Melanesian data using sequential importance sampling on coalescent histories with a proposal distribution of Stephens and Donnelly. The gene trees are drawn to scale with mean ages of mutations and TMRCA calculated as weighted means with likelihood weights on each simulation run. In this method each simulation run is independent. Assuming a 25-year generation time and an effective population size of 20000, the TMRCA estimates in the two trees are 1.08 million and 3.01 million years. Of interest are the mean ages of clades underneath mutations, such as the mutation at site 1358. Fig. 6(a) is constructed with $\theta = 2.55$ and Fig. 6(b) is constructed by assuming that the data are single-nucleotide polymorphism data as an illustration to see the effect. In Fig. 6(b) there is no mutation

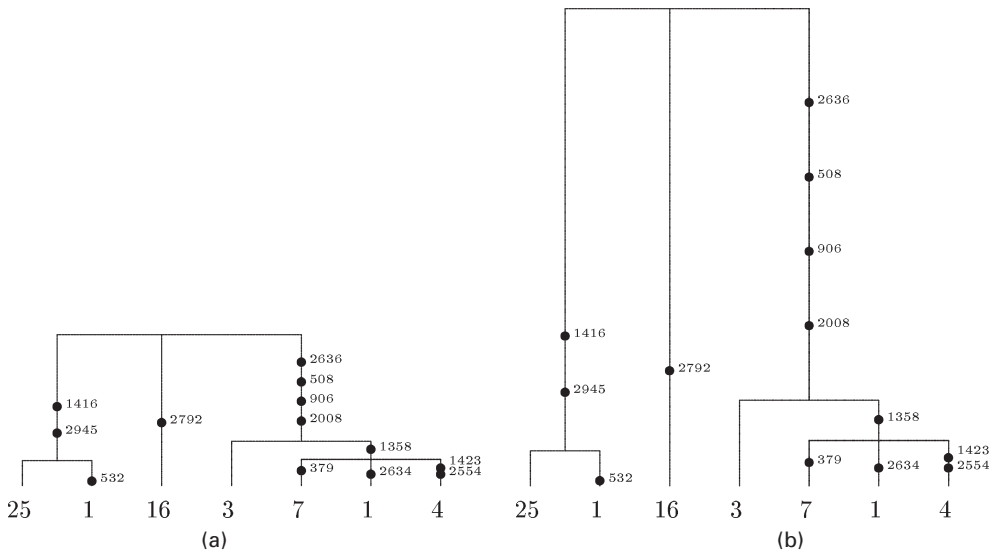

Fig. 6. Melanesian gene trees drawn to scale with expected TMRCA, ages of mutations and ages of clades: in (a) the mutation rate is $\theta = 2.55$ and in (b) the data are assumed to be single-nucleotide polymorphism data and the tree times are computed conditional on the mutant sites segregating, with no assumption about the mutation rate

parameter θ, with the tree being calculated conditional on the segregating sites, supposing that these are the only sites sampled. The tree is approximately three times higher than the tree in Fig. 6(a) because there is no assumption that sites other than those sampled are non-segregating. Lower parts of the tree may be reasonable but an estimate of 3.01 million years seems too high for TMRCA in a biological sense. Although this type of computation can take substantial time, this is a small gene tree and the computing time on a modest personal computer was 260 s with 3 million runs for Fig. 6(a) and 864 s with 10 million runs for Fig. 6(b).

Hilde M. Wilkinson-Herbots (*University College London*)

The interpretation of mitochondrial DNA evidence in court has long been a problem of considerable practical interest, and the method described by Wilson, Weale and Balding is a substantial step forward. However, the mitochondrial DNA 'minisequences' to which their method is applied are mainly useful to exclude suspects quickly and cheaply. Before taking a defendant to court on the basis of mitochondrial DNA evidence, a match for a much longer mitochondrial DNA sequence would normally have been established. For example, the forensic mitochondrial DNA database published by Piercy *et al.* (1993) consists of mitochondrial DNA sequences of approximately 800 nucleotide sites in length. If the authors' method can be readily applied to such 'long' sequences, then their work is of significant practical interest indeed. If however, the analysis of a substantial set of such long sequences proves computationally too demanding at present, then it may be possible to reduce the computational complexity of the problem by taking account of information about the genealogical structure of the mitochondrial DNA gene pool obtained by other methods. Various researchers have studied the major haplogroups that are present in the UK white Caucasian population or the European population as a whole (see for example Wilkinson-Herbots *et al.* (1996) and references therein). Each haplogroup corresponds to a major branch of the genealogical tree and its frequency can be estimated directly from relevant mitochondrial DNA databases. I plan to investigate whether it is possible to use a modified version of the authors' method to estimate the match probability of any individual mitochondrial DNA haplotype, given the frequency of the haplogroup to which it belongs, and focusing primarily on the reduced data set for the haplogroup concerned (although correlations between haplogroups may cause additional difficulties).

Another point where it would be useful to take into account findings from other studies concerns the relative mutation rates of the different nucleotide sites that are included in the minisequence. Whereas part of the polymorphism at the mitochondrial DNA minisequence is due to a few stable, ancient mutations (three of the single-nucleotide polymorphisms characterize major branches of the genealogical tree for the UK white Caucasian mitochondrial gene pool), six of the single-nucleotide polymorphisms included in the minisequence are known to have very high mutation rates (see Wilkinson-Herbots *et al.* (1996) and references therein for evidence at some of these sites). If the authors' method is to be used to evaluate mitochondrial DNA evidence in court cases, then it is important to take this known mutation rate heterogeneity into account, as it may affect the estimates of the match probabilities of uncommon haplotypes.

Mark A. Beaumont (*University of Reading*)

This study provides a significant advance on the original ground breaking paper by Wilson and Balding (1998) and is currently the only approach to allow Bayesian inference of parameters in a model with both population structure and population growth. Methods developed by Wakeley (1999) and Wakeley *et al.* (2001) allow for likelihood-based inference with population structure and growth, and highlight the need to model both aspects jointly. In addition to the effects of population structure there are other phenomena that have the potential to vitiate conclusions drawn about historical changes in population size. These include the effects of ascertainment (where polymorphic loci are deliberately chosen) (Beaumont, 1999; Wakeley *et al.*, 2001), selection at linked sites and the effects of initial population contractions followed by growth (Calmet, 2003). The potential effect of these on inferences from Y-chromosome data are reviewed in Beaumont (2003).

In genealogical modelling there is a general problem of non-identifiability of parameters in the likelihood, which has traditionally been avoided through the use of scaled parameters such as θ. An important innovation in Tavaré *et al.* (1997) and Wilson and Balding (1998) was the use of background information on mutation rates and population sizes to allow for inference on all the parameters of interest. However, it seems to me that only on mutation rates are there grounds to use strongly informative priors. The current and ancestral population sizes and growth rates in the model are unlikely to bear any relation to any estimate of current of historical population size because of their sensitivity to historical metapopulation structure (Wakeley, 2001), the intricate details of which we cannot hope to include directly in our models. In models of population growth, even with proper priors for the mutation rate, improper priors on other

parameters lead to improper posterior distributions (Beaumont, 1999). Therefore, even though we know little about N in the demographic models, it is necessary to impose some limitation on population sizes, and there is a temptation for this to be motivated by the need to obtain good convergence of the Markov chain Monte Carlo algorithm. Yet inferences about changes in population size are sensitive to the priors for N. For example, with the model described in Storz and Beaumont (2002) and Storz *et al.* (2002), if the prior assumes current and ancestral population sizes that are too high the posterior distribution tends to support a model of population growth. This sensitivity could be straightforwardly examined by using several different priors. However, given the difficulties in obtaining convergence, this aspect is probably the most weakly developed part of current Bayesian approaches in population genetics. In conclusion, it seems to me that a period of consolidation is needed in which the sensitivity to model assumptions and specification of the priors is evaluated.

The following contributions were received in writing after the meeting.

Stuart J. E. Baird (*University of California, Berkeley*)

Wilson, Weale and Balding develop a flexible class of Metropolis–Hastings algorithms for drawing inferences about population histories and mutation rates from DNA sequence data. Population structure is generalized to allow splitting of an ancestral population into any number of separate panmictic units. The class of models of population structure that are applicable for inferring human population history is a tiny subset of those that are interesting for making inferences about evolution. Motivated by an interest in broader scale evolutionary inference, inspired by earlier work by two of the authors (Wilson and Balding, 1998), and in consultation with Ian Wilson, a complementary set of algorithms has recently been developed (Baird, 2003) which allows generalization over a wider class of models of population structure. The approach achieves this generality with a trade-off against computation time. A Markov chain Monte Carlo simulation is created with a state consisting of a tree of paths through discrete space and time (Fig. 7). Movement is on a two-dimensional stepping-stone lattice. Between discrete opportunities for movement demes are undisturbed by migration events, and so coalescent probabilities can be described following standard coalescent theory. Proposed transition on the chain state can most succinctly be described as a series of dance steps allowing nodes and paths on the tree to be moved in space and time. The transitions are designed such that change in the tree state is localized, bounded by the nodes connected to the part of the tree being moved and consistent with the stepping-stone paradigm. The process of the Markov chain Monte Carlo simulation can be visualized by iterating the chain and sampling the positions of the lineage paths that make up the tree. Animating the resulting snapshots of the state suggests a label for this approach: the dancing trees algorithm.

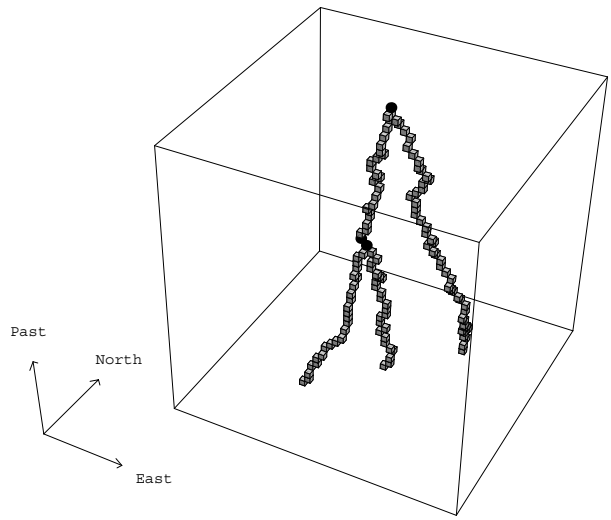

Fig. 7. Example of the explicit state of a genealogy in discrete space–time: cubes represent demes occupied by one lineage; circles represent demes occupied by two lineages

A comparison of inference involving the authors' and dancing trees approaches will be mutually informative. The dancing trees algorithm can be used to define better the set of lineage trees in nature whose history is well approximated by the population splitting model. Conversely this set and others involving islands of panmixis are intrinsically computationally intensive for the dancing trees algorithm. In summary the current work has wide implications: the authors' approach and its complements pave the way towards a sounder understanding of population structure and the evolutionary process.

Martin Lascoux (*Uppsala University*)
In the paper the authors state that 'the present paper is addressed *in part* to statisticians' and, as the paper is being published in the *Journal of the Royal Statistical Society*, this is undoubtedly true. However, I hope that the 'in part' will turn out to be correct as I feel that evolutionary biologists would perhaps benefit most from reading this excellent paper, for which I would like to congratulate the authors. The evolutionary biologists whom I have in mind here are primarily those working with phylogeography, as the coalescent has so far had only a limited influence on this area (at least when humans are not considered). Interestingly, the limited effect that it has had so far can be, at least in part, attributed to precisely the computer program described in the present paper, BATWING, and its predecessor, MICSAT. This is hardly surprising as both programs were tailored for the type of data that are generally produced by phylogeographers, namely variation in non-recombining DNA (chloroplast DNA, mitochondrial DNA, the non-recombining-part of the Y-chromosome). This might also turn out to be one of the main limitations of these programs as only the coalescent analysis of the variation at large numbers of independent loci can lead to more precise inferences for the demographic parameters.

I have two questions for the authors. First, in the light of what has just been said, could their data augmentation approach be extended to include recombination? Second, as shown recently by Ptak and Przeworski (2002), sampling can have an important effect on inferences on the demographic history of species. Interestingly, in his discussion of the paper by Stephens and Donnelly (2000), Ian Wilson has already pointed out that the design and analysis of surveys of different populations have been somewhat neglected by statisticians and geneticists working on the inference of past demographics from molecular variation. Also, in a paper written almost 10 years ago by the first author with N. Barton (Barton and Wilson, 1995) the stage was set for further studies on modelling the coalescent in continuous environments (isolation-by-distance models for instance), i.e. on how to consider jointly the geographical location of individuals and their genotype. How would the authors proceed to find the best sampling strategy to increase our chances of obtaining good estimates of past demographic parameters, given that the sampling strategy depends precisely on knowing something about these parameters? Does this not imply that non-genetics data should explicitly be included in our models?

Raphaël Leblois and Arnaud Estoup (*Centre de Biologie et de Gestion des Populations, Montferrier–Lez*)
We congratulate the authors on their excellent paper. Their methodology represents an important advance towards the goal of fully likelihood-based methods for analysing complex evolutionary scenarios. The treatment of increasingly complex models raises the problem of the validation of methods and programs. Analytical results for the likelihood of a sample of two genes for various population and mutational models can be obtained to check the accuracy of such complicated algorithms (e.g. Nagylaki (1982) and Rousset (1996)). Another important issue is the robustness of algorithms to violations of both the mutation and the demographic assumptions of the model. Simple generation-by-generation coalescence algorithms allow the simulation without approximation of molecular data under virtually any demographic and mutational model and hence can be used to test the robustness and precision of any inferential method.

Because increasingly more models can now be considered, it is crucial to develop criteria for comparing models rather than relying on inferences, from a given model, that fit our beliefs. Did the authors compute the relative likelihood of their four evolutionary models?

The surprisingly low values obtained here for the time since the most recent common ancestor, TMRCA, raise several questions. To what extent could low TMRCA values reflect inappropriate prior assumptions for the mutation rate of microsatellites? More importantly, the possibility of migration between populations is expected to reduce TMRCA substantially as well as time split estimations in a model with no migration. Moreover, since the possibility of homogenizing selection acting on the Y-chromosome may also explain low TMRCA values, it would be worth performing similar analyses on independent and presumably neutral microsatellite loci on autosomal chromosomes.

Finally, in agreement with Fu and Li (1999), our analysis of particularly complex evolutionary histories indicates that, in such cases, inferential methods that are not fully likelihood based still appear to be the

best option available (e.g. Pritchard *et al.* (1999), Estoup *et al.* (2001) and Beaumont *et al.* (2002)). These methods combine the computational convenience of summary statistics with the advantages of the Bayesian paradigm and can handle complex models provided that the simulation of data under the model is feasible. Simulation results have shown that the computational and statistical efficiency of such methods compares favourably with those of the Markov chain Monte Carlo method described here (Beaumont *et al.*, 2002). However, the Markov chain Monte Carlo based method still appears consistently superior to the summary-statistic-based methods, highlighting that it is well worth making the effort to obtain full data inferences if possible.

Rasmus Nielsen (*Cornell University, Ithaca*) **and Jody Hey** (*Rutgers University, Piscataway*)

Several likelihood-based methods for analysing data from multiple populations have been developed in recent years. The methods differ with respect to population genetic assumptions and with respect to the type of data that they are applicable to. Wilson, Weale and Balding have chosen a demographic model of population splitting with no migration between populations. Other likelihood-based methods for estimating parameters in demographic models with population splitting have been proposed by Nielsen *et al.* (1998), Nielsen (1998) and Nielsen and Slatkin (2000). In contrast, the likelihood methods of Beerli and Felsenstein (1999, 2001) and Bahlo and Griffiths (2000) assume infinite divergence times among populations but allow for arbitrary levels of migration between populations. The method of Nielsen and Wakeley (2001) allows for both finite divergence and migration but is only applicable to pairs of populations. The method of Wilson, Weale and Balding also differs from these methods by incorporating population growth. Although all the models naturally are simplifications of the true model, the question arises which of these models is most appropriate. There are undoubtedly organisms in which the authors' model is adequate; however, in the human genetics community there appears to be growing concern that human evolution cannot be described by using models that ignore migration. Evidence of gene flow between human populations has been found at local scales (see for example Papiha *et al.* (1997), Lum *et al.* (2002) and Fix (1999)), across continents (Bandelt *et al.*, 2001; Sokal *et al.*, 1991), as well as between continents (Hammer *et al.*, 1998).

To illustrate the effect of migration, we reanalysed the β-globin data set of Harding *et al.* (1997) for 46 European and 24 Asian individuals, using the method in Nielsen and Wakeley (2001) which incorporates both population splitting and migration. In Fig. 8 we present the marginal posterior distribution of the scaled migration parameter M, assuming uniform priors for all parameters. Note that very little of the probability mass is located around $M = 0$. The marginal posterior distribution for the splitting time T is a strictly increasing function of T (not shown). The data appear to be compatible with a model of equilibrium migration as in Beerli and Felsenstein (1999, 2001) and Bahlo and Griffiths (2000), but not with a model of population splitting without migration. Although we consider the authors' method a great improvement on previous methods for estimating population splitting times in the absence of migration, methods that incorporate migration may be more applicable to the analysis of human genetics data.

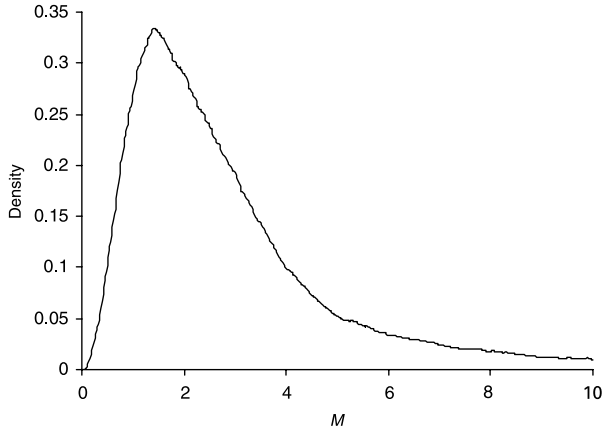

Fig. 8. Marginal posterior distribution for the scaled migration rate M for the β-globin data

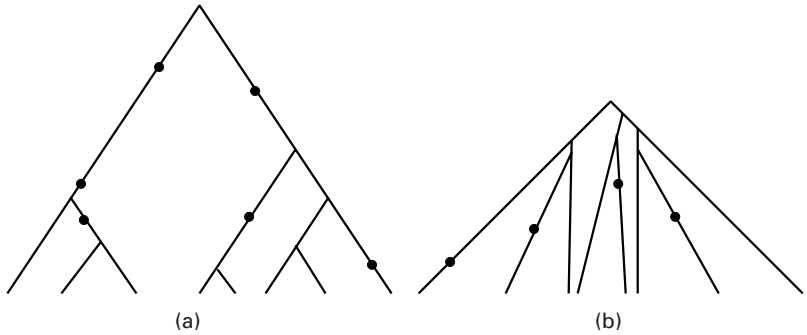

Fig. 9. Genealogy of a sample from a population, assuming (a) constant population size or (b) population growth (note that evolution under constant population size is expected to result in more haplotypes at intermediate frequencies than under a growing population scenario): •, mutations

Michael P. H. Stumpf and Hilde M. Wilkinson-Herbots (*University College London*)
In Section 5.4.2, the authors comment that no appreciable differences in the match probabilities of the mitochondrial DNA minisequences were found when they used the coalescent model with population growth (for which they do not show their results) rather than the standard coalescent model with constant population size. This is surprising, as these two demographic scenarios are expected to lead to different haplotype frequency distributions. Genealogies of populations that have experienced growth tend to be star like. If the mutation rate is not too high compared with the timescale that is involved, this is expected to lead to a very common haplotype and some rare haplotypes, as is illustrated in Fig. 9(b).

In a constant-sized population, by contrast, we expect to obtain more haplotypes at moderate frequencies, giving a different haplotype frequency distribution (see Fig. 9(a)). If a coalescent model with constant population size is used as a prior when estimating match probabilities of haplotypes that actually evolved in an expanding population, we would therefore expect the match probability of the most common haplotype to be underestimated; conversely match probabilities of rare haplotypes may be overestimated. This may at least in part explain the low estimates obtained by the authors (Table 11) for the match probability of the most common haplotype in each of the three ethnic groups considered, compared with the 'naïve' estimate (which is the observed relative frequency of the haplotype)—for a common haplotype we would expect the latter estimate to be reasonably accurate. An inappropriate use of the coalescent model with constant population size might also explain the relatively high estimates that the authors obtained for the match probabilities of the 'similar' and the 'dissimilar' haplotypes that are listed in Table 11 for the Caucasian population, where population growth is believed to have been particularly strong (see also Section 5.2.1.3).

We have verified the above-described effect of population growth on the expected haplotype frequencies by a large number of coalescent simulations under constant population size *versus* population growth (our results are not shown).

The **authors** replied later, in writing, as follows.

We are gratified by the positive and constructive contributions, and we thank all the discussants for their comments. There is considerable overlap of the questions and comments, and we focus on issues raised by several discussants.

More complex demographic models
Geographical structuring of, and migration between, populations is thought to underlie many observed patterns in human DNA data. Lascoux mentions the importance of sampling strategies: our allowance of different population sizes gives some flexibility to incorporate these effects. Nielsen and Hey report evidence of migration in a superset of the H97 data, Leblois and Estoup suggest that ignoring migration may have reduced our TMRCA estimates, and several other contributors mention the desirability of modelling migration within BATWING. We agree. However, the problem of the number of migration parameters rising quadratically with the number of subpopulations (Section 3.1.3) would have been substantial for the 13-subpopulation data set, and an assumption of a common migration parameter would have been suspect. Our splitting model for population structure is unrealistic in some respects but captures some principal aspects of structured data while being computationally relatively unburdensome.

Although computer-intensive methods such as Markov chain Monte Carlo (MCMC) methods permit the analysis of complex models, often generality of model structure is acquired at the expense of efficiency of the algorithm. Inevitably our choice of models was restricted by the computer power that was available at the time: migration would have considerably increased the computational burden. With further increases in computer power, this approach may now be feasible and we shall make efforts to incorporate migration in a future version of BATWING.

Lascoux and Baird want to draw inferences about phylogeography—using the joint information about location and genotype to draw inferences about the processes of evolution. These models are very much more complex. Baird gives an account of his methodology—the coalescent on a two-dimensional grid. He employs auxiliary variables liberally, to try to reduce the computations required for each change in the tree. We commend the approach but note that many auxiliary variables can lead to more problems with mixing.

Yang mentions the similarity of our model to a model for the estimation of ancestral population sizes in closely related species. Our population supertree model is well suited to this application, and it would be interesting to compare results.

Parameterization

Several contributors mention the confounding of effective population size N_e with rate parameters such as μ. This reflects the fact that, if we know only that k events have occurred in an unknown time period, we cannot distinguish a low rate–large time scenario from high rate–small time. For this reason population geneticists have traditionally been limited to working only with $\theta = 2N_e\mu$, a severe limitation since the timescale parameter N_e is required to convert coalescent time units into practically useful units such as generations or years.

Drummond *et al.* (2002) overcame the confounding problem by using time-stamped data, a possibility that is also mentioned by Stephens. Time-stamped data are rarely available for humans, although archaeological evidence can give some help.

BATWING allows users to work either with θ or with N_e and μ separately. As noted by Beaumont, this constitutes a major advance, but it carries the inevitable consequence of sensitivity to the prior: the data are informative about θ, but the 'allocation' of this information between N and μ depends entirely on the prior. Although we agree that there are difficulties with interpreting the available information, there is nevertheless substantial background information about both N_e and μ. Our approach has been to use this information as best we can, making explicit our choice of priors and the evidence on which they are based.

Beaumont and Drummond and Nicholls seem happy with our choice of prior for μ but mention possible problems with priors for N. Our gamma prior is reasonably diffuse, but we agree that it could have some influence on growth rate estimates. We distribute a program with BATWING that allows users to simulate from their prior to explore some of its implications, e.g. about TMRCA.

BATWING works with identifiable parameters internally, so non-identifiable parameters do not cause problems with mixing. Stephens raises the related problem of 'weakly identifiable' parameters: a subset of the parameters such that changes in some members of the subset can be largely compensated (so that the likelihood is almost unchanged) by changes in other members. Weakly identifiable parameters are both sensitive to prior assumptions and potentially problematic for mixing. We highlight one case in the paper: the growth rate and time since the start of growth. Since we work with complex models, involving typically hundreds of parameters, sets of weakly identifiable parameters are practically inevitable, and it is infeasible to diagnose them all. Our approach has been to formulate priors as carefully as possible, and to check mixing as much as we can.

Combining information across loci may help with some cases of weak identifiability. However, Dawson makes the point that selection can affect N_e, and that we should be careful about using the same N_e-values for different loci, even allowing for the difference in the number of chromosomes for Y and nuclear DNA.

Efficiency issues

Stephens asks how our inferential methods compare with other strategies. The importance sampling method of Stephens and Donnelly (2000) works well with one or two linked short tandem repeats but performs less well when there are many linked short tandem repeat loci. We agree with Leblois and Estoup that, because fully likelihood-based methods remain in a phase of development, non-likelihood methods may still be the best option in many settings. The rejection sampling methods that they highlight, briefly described in Section 6, have advantages and there is scope for improving these methods (Beaumont *et al.*, 2002). However, there are no general principles for finding the good summary statistics that are needed, or for assessing the resulting approximation. We also concur with Leblois and Estoup that likelihood-based methods are preferred when available.

Our experience does not accord with Drummond and Nicholls's view that peeling is preferable to the use of auxiliary variables for highly variable data sets. Early versions of BATWING used a peeling algorithm, but the auxiliary variable approach was found to be much better for highly variable short tandem repeat data. We agree that peeling may be preferable for sequence data with low mutation rates, since mixing of auxiliary variables may then be poor. Standard auxiliary variable approaches are unlikely to be efficient for monomorphic sites, but these can be treated separately. We have found that our auxiliary variable approach works well at least for the H97 sequence data set, and computation time is linear in sequence length for both approaches.

Stephens is concerned about model checking, which was not a focus of our paper. We expect that our most general model (splitting with growth) is substantially superior to simpler models that are in widespread use, yet it is still inadequate to capture all important features of the data. We have reported informal model validation via comparison across our models and with the results of other researchers. Nevertheless we agree that quantitative model comparison and assessment are a priority for the future. Under the standard coalescent, increasing the sample size often has little effect of inferences; thus there may be little loss in reserving some data to be used for model testing only, not model fitting.

Mitochondrial DNA match probabilities

Stumpf and Wilkinson-Herbots suggest that a high growth rate should influence the match probability in forensic inference on mitochondrial DNA data. We concur that evidence for growth has been reported from mitochondrial DNA data sets (Excoffier, 2002), and this should indeed have some effect on match probabilities. However, the effect of growth on inference from an observed data set may be much less than its effect on simulations that are not constrained by data.

Wilkinson-Herbots notes that much longer mitochondrial DNA sequences would nowadays be routinely typed, and also that mutation rate heterogeneity is found in such sequences. We explored the effect of a variable mutation rate in a simple way by allowing for different rates at each single-nucleotide polymorphism and also by splitting single-nucleotide polymorphisms into 'high' and 'low' rate categories according to published evidence, but we admit that a more sophisticated rate heterogeneity model would be preferable. It will be worth investigating the proposal of Wilkinson-Herbots to perform BATWING within subclades only (we suggest that fitting an exponential growth model will compensate in part for the different genealogical structure of subclades relative to the whole tree). Because of the relatively high mutation rate of the mitochondrial DNA control region, this would provide an interesting test of the relative merits of peeling algorithms and auxiliary variable approaches.

We do not propose our algorithm for the routine calculation of mitochondrial DNA match probabilities, because of the remaining questions about model validity and because of the computation time that is required for each calculation. Instead, our goals in undertaking the mitochondrial DNA match probability were

(a) to indicate some of the possibilities that genealogical modelling opens and
(b) to check the validity of the naïve estimator against a more sophisticated approach in at least some settings.

We found that, although not providing a bound, the naïve estimator is likely to be adequate in practice for the scenarios that we considered.

References in the discussion

Bahlo, M. and Griffiths, R. C. (2000) Inference from gene trees in a subdivided population. *Theoret. Popln Biol.*, **57**, 79–95.

Baird, S. J. E. (2003) A novel approximation to the spatial coalescent: the dancing trees algorithm . To be published.

Bandelt, H. J., Alves-Silva, J., Guimaraes, P. E., Santos, M. S., Brehm, A., Pereira, L., Coppa, A., Larruga, J. M., Rengo, C., Scozzari, R., Torroni, A., Prata, M. J., Amorim, A., Prado, V. F. and Pena, S. D. (2001) Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade. *Ann. Hum. Genet.*, **65**, 549–563.

Barton, N. H. (1998) The effect of hitch-hiking on neutral genealogies. *Genet. Res.*, **72**, 123–133.

Barton, N. H. and Wilson, I. (1995) Genealogies and geography. *Proc. R. Soc. Lond.* B, **349**, 49–59.

Beaumont, M. A. (1999) Detecting population expansion and decline using microsatellites. *Genetics*, **153**, 2013–2029.

Beaumont, M. A. (2003) Recent developments in genetic data analysis: what can they tell us about human demographic history? *Heredity*, to be published.

Beaumont, M. and Nichols, R. A. (1996) Evaluating loci for use in the genetic analysis of population structure. *Genetics*, **153**, 2013–2029.
Beaumont, M. A., Zhang, W. and Balding, D. J. (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.
Beerli, P. and Felsenstein, J. (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, **152**, 763–773.
Beerli, P. and Felsenstein, J. (2001) Maximum-likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc. Natn. Acad. Sci. USA*, **98**, 4563–4568.
Bowcock, A. M., Kidd, J. R., Mountain, J. L., Hebert, J. M., Carotenuto, L., Kidd, K. K. and Cavalli-Sforza, L. L. (1991) Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. *Genetics*, **88**, 839–843.
Calmet, C. (2003) Inférences sur l'histoire des populations à partir de leur diversité genetique: étude de séquences démographiques de type fondation-explosion. *PhD Thesis*. University of Paris, Paris.
Drummond, A. J., Nicholls, G. K., Rodrigo, A. G. and Solomon, W. (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, **161**, 1307–1320.
Edwards, A. W. F. (1970) Estimation of the branch points of a branching diffusion process (with discussion). *J. R. Statist. Soc.* B, **32**, 155–174.
Estoup, A., Wilson, I. J., Sullivan, C., Cornuet, J.-M. and Moritz, C. (2001) Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. *Genetics*, **159**, 1671–1687.
Excoffier, L. (2002) Human demographic history: refining the recent African origin model. *Curr. Opin. Genet. Devlpmnt*, **12**, 675–682.
Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Molec. Evoln*, **17**, 368–376.
Fix, A. G. (1999) *Migration and Colonization in Human Microevolution*. Cambridge: Cambridge University Press.
Fu, Y. X. and Li, W. H. (1999) Coalescing into the 21st century: an overview and prospects of coalescent theory. *Theoret. Popln Biol.*, **56**, 1–10.
Griffiths, R. C. and Tavaré, S. (1994) Sampling theory for neutral alleles in a varying environment. *Phil. Trans. R. Soc. Lond.* B, **344**, 403–410.
Griffiths, R. C. and Tavaré, S. (1999) The age of mutations in gene trees. *Ann. Appl. Probab.*, **9**, 567–590.
Hammer, M. F., Karafet, T., Rasanayagam, A., Wood, E. T., Altheide, T. K., Jenkins, T., Griffiths, R. C., Templeton, A. R. and Zegura, S. L. (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Molec. Biol. Evoln*, **15**, 427–441.
Harding, R. M., Fullerton, S. M., Griffiths, R. C. and Clegg, J. B. (1997) A gene tree for β-globin sequences from Melanesia. *J. Molec. Evoln*, **44**, suppl. 1, S133–S138.
Kuhner, M. K., Yamato, J. and Felsenstein, J. (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, **140**, 1421–1430.
Lum, J. K., Jorde, L. B. and Schiefenhovel, W. (2002) Affinities among Melanesians, Micronesians, and Polynesians: a neutral biparental genetic perspective. *Hum. Biol.*, **74**, 413–430.
Nagylaki, T. (1982) Geographical invariance in population genetics. *Theoret. Popln Biol.*, **99**, 159–172.
Nielsen, R. (1998) Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. *Theoret. Popln Biol.*, **53**, 143–151.
Nielsen, R., Mountain, J. L., Huelsenbeck, J. P. and Slatkin, M. (1998) Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution*, **52**, 669–677.
Nielsen, R. and Slatkin, M. (2000) Analysis of population subdivision using di-allelic models. *Evolution*, **54**, 44–50.
Nielsen, R. and Wakeley, J. (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.
Papiha, S. S., Singh, B. N., Lanchbury, J. S., Mastana, S. S. and Rao, Y. S. (1997) Genetic study of the tribal populations of Andhra Pradesh, south India. *Hum. Biol.*, **69**, 171–199.
Piercy, R., Sullivan, K. M., Benson, N. and Gill, P. (1993) The application of mitochondrial DNA typing to the study of white Caucasian genetic identification. *Int. J. Leg. Med.*, **106**, 85–90.
Pritchard, J. K., Seielstad, M. T., Prez-Lezaum, A. and Feldman, M. W. (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molec. Biol. Evoln*, **16**, 1791–1798.
Ptak, S. E. and Przeworski, M. (2002) Evidence for population growth in humans is confounded by fine-scale population structure. *Trends Genet.*, **18**, 559–563.
Rannala, B. and Yang, Z. (2003) Bayes estimation of species divergence times and ancestral population sizes using multi-locus DNA sequences. Submitted to *Genetics*.
Rousset, F. (1996) Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics*, **142**, 1357–1362.
Sokal, R. R., Oden, N. L. and Wilson, C. (1991) Genetic evidence for the spread of agriculture in Europe by demic diffusion. *Nature*, **351**, 143–145.

Stephens, M. and Donnelly, P. (2000) Inference in molecular population genetics (with discussion). *J. R. Statist. Soc.* B, **62**, 605–655.
Storz, J. F. and Beaumont, M. A. (2002) Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution*, **56**, 154–166.
Storz, J. F., Beaumont, M. A. and Alberts, S. C. (2002) Genetic evidence for long-term population decline in a savannah-dwelling primate: inferences from a hierarchical Bayesian model. *Molec. Biol. Evoln*, **19**, 1981–1990.
Takahata, N., Satta, Y. and Klein, J. (1995) Divergence time and population size in the lineage leading to modern humans. *Theoret. Popln Biol.*, **48**, 198–221.
Tavaré, S., Balding, D. J., Griffiths, R. C. and Donnelly, P. (1997) Inferring coalescence times from DNA sequence data. *Genetics*, **145**, 505–518.
Vitalis, R., Dawson, K. J. and Boursot, P. (2001) Interpretation of variation across marker loci as evidence of selection. *Genetics*, **158**, 1811–1823.
Wakeley, J. (1999) Nonequilibrium migration in human history. *Genetics*, **153**, 1863–1871.
Wakeley, J. (2001) The coalescent in an island model of population subdivision with variation among demes. *Theoret. Popln Biol.*, **59**, 133–144.
Wakeley, J., Nielsen, R., Liu-Cordero, S. N. and Ardlie, K. (2001) The discovery of single-nucleotide polymorphisms—and inferences about human demographic history. *Am. J. Hum. Genet.*, **69**, 1332–1347.
Wilkinson-Herbots, H. M., Richards, M. B., Forster, P. and Sykes, B. C. (1996) Site 73 in hypervariable region II of the human mitochondrial genome and the origin of European populations. *Ann. Hum. Genet.*, **60**, 499–508.
Wilson, I. J. and Balding, D. J. (1998) Genealogical inference from microsatellite data. *Genetics*, **150**, 499–510.
Yang, Z. (2002) Likelihood and Bayes estimation of ancestral population sizes in Hominoids using data from multiple loci. *Genetics*, **162**, 1811–1823.
Yang, Z., Kumar, S. and Nei, M. (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, **141**, 1641–1650.