Modern computational approaches for analysing molecular genetic variation data

Paul Marjoram and Simon Tavaré*‡*

Abstract | An explosive growth is occurring in the quantity, quality and complexity of molecular variation data that are being collected. Historically, such data have been analysed by using model-based methods. Models are useful for sharpening intuition, for explanation and for prediction: they add to our understanding of how the data were formed, and they can provide quantitative answers to questions of interest. We outline some of these model-based approaches, including the coalescent, and discuss the applicability of the computational methods that are necessary given the highly complex nature of current and future data sets.

There has never been a better time to analyse molecular variation data from natural populations. We are in the midst of an explosive growth in both the amount of molecular data being generated and the computational power available to analyse them. An increasing variety of computational methodologies are now available to aid in analysing and interpreting such data. However, the speed with which the field is changing means that previously useful methods will be less successful in future. As such, it is perhaps time to take a look at where the field is, where it is heading, and to contemplate the ways in which computational methodologies are changing to meet the challenges of current and forthcoming data. In particular, we focus on the move from exact to more approximate methods — that is, on the growing need to use simplified models or summaries of the data.

It is impossible to survey all the applications of such methods. Consequently, we have chosen to concentrate on the area that has seen perhaps the greatest number of applications: population genetics and the methods that have been developed to answer the questions that arise within species, rather than between species.

One of the aims of population genetics is to understand the forces that shape patterns of molecular genetic variation. Over the past 40 years, this variation has been assayed in different ways. The first method identified electrophoretic variants[1]. Restriction fragment length polymorphisms (RFLPs) soon followed[2,3]. Subsequently, DNA sequence variants were identified in *Drosophila melanogaster*[4], and these studies were followed by surveys of sequence variation in human mitochondrial DNA[5,6] and Y chromosomes[7,8]. Variation at autosomal loci in larger populations is often based on microsatellite marker loci[9,10]. More recently, the advent of fast sequencing and genotyping technologies has made the collection of large data sets of genetic variation a reality in various organisms[11–13].

From its inception, theoretical population genetics has had strong quantitative underpinnings[14,15]. From a methodological perspective, the focus of this field has been to develop detailed stochastic models to describe the evolution of allele frequencies over time at particular loci. A model is typically a relatively simple mathematical formulation of the biological processes that produce our data. A model incorporates parameters of interest, such as mutation or recombination rates. All models that are discussed here are stochastic: there is no predetermined outcome, but instead many outcomes are possible. Traditionally, models in theoretical population genetics have allowed researchers to predict how patterns of variation would be affected by forces such as genetic drift, selection, migration and recombination. Although understanding these models in a 'pre-computational' environment led to several interesting developments in probability, such as the development of coalescent theory (see later discussion), the modern approach is to make intensive use of simulation methods. This approach is largely motivated by the current rapid growth in computational power and the concurrent increase in the quantity and complexity of data that are being collected.

Restriction fragment length polymorphisms
Variations between individuals in the lengths of DNA regions that are cut by a particular endonuclease.

Microsatellite marker loci
Polymorphic loci at which short DNA sequences are repeated a varying number of times.

*University of Southern California, Keck School of Medicine, Preventive Medicine, 1540 Alcazar Street, CHP-220, Los Angeles, California 90089-99011, USA.
‡Program in Molecular and Computational Biology, University of Southern California, 1050 Childs Way, Los Angeles, California 90089-2910, USA, and Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WA, UK.
Correspondence to S.T.
e-mail: stavare@usc.edu

doi:10.1038/nrg1961

Box 1 | Successful applications of model-based approaches

Here we highlight some successful ongoing applications of the model-based approaches. We also give some representative references of a computational nature.

Mutation and recombination rates

An early focus of model-based approaches was to estimate population parameters such as mutation and recombination rates. Using computational methods such as those reviewed in this paper, Griffiths and Tavaré[38] and Kuhner *et al.*[79] developed estimators of mutation rate. A wide variety of estimators have been developed to estimate the recombination rate. Some of these are surveyed in REFS 80,81, but we highlight one or two here. In particular we draw attention to the 'composite-likelihood' estimators[82,83], which use approximate methods similar to the approaches that are discussed in the section on approximate Bayesian computation.

Demographic parameters

There has also been great interest in model-based inference relating to demography. For example, Beerli and Felsenstein developed a model-based procedure for estimating migration rates[84]; Griffiths and Tavaré[39] and Kuhner *et al.*[85] developed methods for identifying population-size fluctuations; whereas Pritchard *et al.*[86] introduced the popular Structure software for identifying population substructure and assigning samples to subpopulations.

Selection

Another ongoing focus of research has been the development of methods for the discovery of regions of the genome that are under selective pressure. For example, Voight *et al.*[87] identified widespread evidence for recent selective events in the HapMap data from the International HapMap Project. Pollinger *et al.*[88] used a model-based approach to identify selective sweeps in dogs.

Ancestral inference

Historically, there has been interest in inferring the time to the most recent common ancestor and the age of specific mutations. The most famous example of the first is the large body of literature regarding the identity, location and age of mitochondrial Eve[5,32]. Examples of the second can be found in REFS 30,31.

Structure of the genome

The HapMap project[12] has led to an increasing effort to understand the structure of the genome. For example, Nordborg and Tavaré[89] surveyed the behaviour of linkage disequilibrium in the human genome.

As the length of the chromosomal region for which data are collected grows, the task of reconstructing haplotypes from SNP data becomes more difficult. Perhaps the most popular tool for this is the PHASE software, which was first introduced by Stephens *et al.*[90] and refined by Scheet and Stephens[91] as the recently released fastPHASE.

Another particular focus of present-day research is the identification of recombination hot spots. There are many recent and ongoing projects, of which we mention a representative few: Crawford *et al.*[92] investigated the pattern of fine-scale recombination-rate variation in the human genome and found widespread variation; McVean *et al.*[83] used a Markov chain Monte Carlo scheme to estimate recombination-rate variation; Fearnhead and Smith[93] derived an approximate method for estimating recombination rate and use it to detect hot spots; Li and Stephens[65] used an alternative, approximate approach to the same problem; Myers *et al.*[94] identified over 25,000 recombination hot spots, genome-wide; and Tiemann-Boege *et al.*[95] used approximate Bayesian computation to estimate recombination rates from sperm-typing data.

Human association studies

An important area of interest is that of association studies for mapping disease genes (which are discussed in the review by Balding in this issue[96]).

Stochastic model
A model that is used to describe the behaviour of a random process.

Coalescent
A popular probabilistic model for the evolution of 'individuals'. Individuals might be single nucleotides, mitochondrial DNA, chromosomes and so on, depending on the context.

There are two different, but related, uses of the word 'simulation' in this context. The first involves simulating the data under a stochastic model, thereby producing data sets that are representative outcomes of the evolutionary process; data sets that result from the same model might differ because of the effects of chance. For example, this approach might be used to examine the degree of variability that is possible in data that have been produced under a proposed model of human evolution. Do independent runs of the evolutionary scenario result in data with similar features, or do data vary substantially between replicates[16]? The second sense in which we use the word simulation refers to the use of simulation-based methods of statistical inference to estimate parameters, such as mutation or recombination rates, from a particular example of the evolutionary process that is described by the model. Here we start with a single, observed data set and use simulation of data under a variety of parameter values in an attempt to infer the relative likelihood of particular parameter values, given the data.

In this review we discuss the main model-based methods that can be applied to large population genetic data sets of the types that are alluded to above. These methods involve an interplay between the two uses of simulation that we have described. Some recent successful applications of these methods are shown in BOX 1.

We begin by introducing the most common population genetics model, the coalescent. In brief, the coalescent provides a theoretical description of the ancestral relationships that exist among a sample of chromosomal segments (such as DNA sequences) taken from a population. Its particular merit is that it ignores lineages that do not appear in the sample and therefore provides for the efficient simulation of data. To understand how models such as the coalescent can be used in a statistical analysis, we need to outline the general approach of model-based analysis; we give an example of the use of these methods by estimating the mutation rate and time to 'mitochondrial Eve' (mtEve). We then describe in some detail the modern simulation-based methods that have exploited a model such as the coalescent to infer demographic parameters. We conclude with a discussion of the present and future developments in statistical modelling. The paper will be useful to the non-specialist, in that the particular focus of the review is but one example of the many areas in which the quantity and complexity of data is rapidly increasing. The methods and developments we discuss below have parallels that are widely applicable in the genetics community.

The coalescent: a population genetics model

As discussed in the introduction, stochastic models have had an important role in population genetics for many years. Simulating models under varying scenarios (that is, parameter values) allows us to explore the effect that changing those parameters has on the data that might typically be observed. As computational power has improved, models have grown more complex, and have therefore become more realistic. Nonetheless, a model must be simple enough to be computationally tractable. For many years, the coalescent[17–19] has been the basic stochastic model in the analysis of genetic variation data that have been obtained from population samples.

Basic features of the coalescent. The coalescent provides a description of the genealogical relationships among a random sample of DNA fragments, and it provides a way to simulate samples of such fragments under many genetic and demographic scenarios. Rather than simulate the evolution of a sample forwards in time, the coalescent models the evolutionary process by going backwards along the lineages that gave rise to that sample to identify points at which pairs of fragments join (that is, coalesce) at a common ancestor fragment. We provide a brief description of the coalescent in BOX 2, together with some useful reviews of coalescent theory that provide a deeper introduction than space allows here.

The simplest form of the coalescent occurs when fragments inherit their genetic material from only one parental fragment, without recombination (BOX 2a–c). The canonical example is mitochondrial DNA. In such a setting, the topology of the ancestry is a tree. There are two key parameters. The first is the rate at which the lines of ancestry within the tree coalesce. This depends on the probability that two fragments have the same parental fragment in the previous generation; the rate is inversely proportional to the population size. However, for mathematical convenience, the standard formulation of the coalescent is run on a transformed timescale in which, on average, for any pair of fragments, there is one coalescence event for each unit of time. In this article we cite all times in coalescent units to avoid arcane discussions about population sizes. The second parameter reflects the rate at which mutations occur.

Several factors complicate the structure of the coalescent in more realistic settings; these include demography, recombination, gene conversion and selection. A full discussion of these is outside the scope of this article (we refer readers to REFS 20–22). Instead, we focus on the effects of recombination, which are illustrated in BOX 2d. Recombination events cause lines of ancestry to bifurcate as we move up the page (back in time). A new parameter is introduced to reflect the rate at which this occurs. As indicated in BOX 2d, the coalescent topology is now a graph rather than a tree, but it remains the case that the ancestry of any particular position along the DNA fragment can be described by a coalescent tree. The trees that apply at different positions are correlated, with the extent of the correlation decreasing as the distance between the two positions increases. It is this correlation that induces linkage disequilibrium, the non-random association of alleles at different positions along the fragment.

The standard tool for simulating the coalescent is Hudson's ms[23]. This program simulates typical data sets that result from the coalescent model with user-specified parameter values. It allows for features such as population growth, subdivision and admixture. Many other programs for simulating data, such as the SIMCOAL program (developed by Excoffier *et al.*, see online links), have been devised to deal with more complicated demographic models, for example, or for rapid simulation[24,25]. Not everything can be simulated backwards through the coalescent, particularly some forms of selection. This limitation has prompted a return to forward simulation — exemplified by simuPOP[26] and by the FPG program of Hey *et al.* (see online links) — which has been made feasible by the recent marked increases in computational power.

Applications of the coalescent. The coalescent has traditionally been used in several ways. At its simplest, it is used as a simulation tool. It provides a concise, efficient way to simulate multiple data sets under plausible evolutionary scenarios (that is, to simulate data in the first sense that was defined in the introduction). Data that are simulated using the coalescent are also used to underpin methods for statistical analysis, such as those we discuss later. The main limitations of the coalescent are, for example, that it assumes rather simplistic models for population structure and selection, and can become highly computationally intensive when simulating long chromosomal regions[24].

One of the classical statistical problems in population genetics has been to estimate population parameters such as mutation, migration, recombination and growth rates. This has usually been approached in a classical statistical style, by treating the data as though they were generated by a suitable stochastic model (such as the coalescent), and estimating the parameters of this model. To do this, the traditional statistical paradigm of maximum likelihood has often been used, classical examples being the celebrated results of Ewens[27] in estimating the mutation rate from electrophoretic data, and Watterson[28], concerning the estimation of parameters from DNA sequence data. The ability to generate ever-richer snapshots of variation soon revealed the problem: formal statistical methods had to become much more computationally orientated.

Another common problem in the literature concerns tests of neutrality: do a particular set of gene frequencies correspond to what would be expected under neutral evolution? A common approach is to devise a test statistic with a distribution that is sensitive to departures from neutrality, and then find (either explicitly or by simulation) the distribution of the statistic under neutrality, assuming a particular model for the evolution of the data. The classical example of this for DNA sequence data is the collection of tests that are based around Tajima's *D*[29].

The last class of statistical problems addressed in the population genetics literature concerns issues such as the estimation of the age of a mutation[30,31], or the time to the most recent common ancestor (TMRCA), of a set of sequences[5,32]. Although several approaches have been used to address such questions, the typical strategy has been to calculate the probability distribution of the age, conditional on the observed data.

Examples of model-based analysis

We now introduce simple examples to explain model-based analysis techniques. These examples will be used throughout the paper for this purpose. The first example is inference concerning the TMRCA of a given sample of sequences. The concept of the TMRCA is introduced in BOX 2. The second example concerns estimation of the mutation rate. We use mitochondrial data in our examples.

In a model-based analysis, we have a model that reflects, to an acceptable degree of accuracy, how the data were generated. The behaviour of the model is determined by the values of a set of parameters. We then

Selective sweep
The increase in the frequency of an allele (and closely linked chromosomal segments) that is caused by selection for the allele. Sweeps initially reduce variation and subsequently lead to increased homozygosity.

Likelihood
The probability of the data under a particular model, viewed as a function of the parameters of that model (note that data discussed in this paper are discrete).

Mitochondrial Eve
The most recent maternal common ancestor of the entire human mitochondrial population.

Gene conversion
A non-reciprocal recombination process that results in the alteration of the sequence of a gene to that of its homologue during meiosis.

Admixture
Gene flow between differentiated populations.

Maximum likelihood
A statistical analysis in which one aims to find the parameter value that maximizes the likelihood of the data.

Test statistic
A numerical summary of the data that is used to measure support for a null hypothesis. Either the test statistic has a known probability distribution (such as χ^2) under the null hypothesis, or its null distribution is approximated computationally.

Tajima's *D*
A statistic that compares the observed nucleotide diversity to what is expected under a neutral, constant population-sized model.

Box 2 | The coalescent

Here we introduce the most popular population genetics model: the coalescent. We begin by introducing the simplest form, in which there is no recombination, and then discuss the version that applies in a more realistic setting.

Coalescent without recombination

Panels **a**–**c** illustrate the intuition that underlies the coalescent using a population of DNA fragments that are evolving according to a Wright–Fisher model — that is, in the absence of recombination, in a population of constant size.

Panel **a** shows a schematic of an evolving population. In this simplified representation of evolution, each row corresponds to a single generation, and each blue circle denotes a fragment in that generation. Generations are replaced in their entirety by their offspring, with arrows running from the parental fragment to the offspring fragment. The present day is represented by the bottom row, with each higher row representing one generation further back into the past.

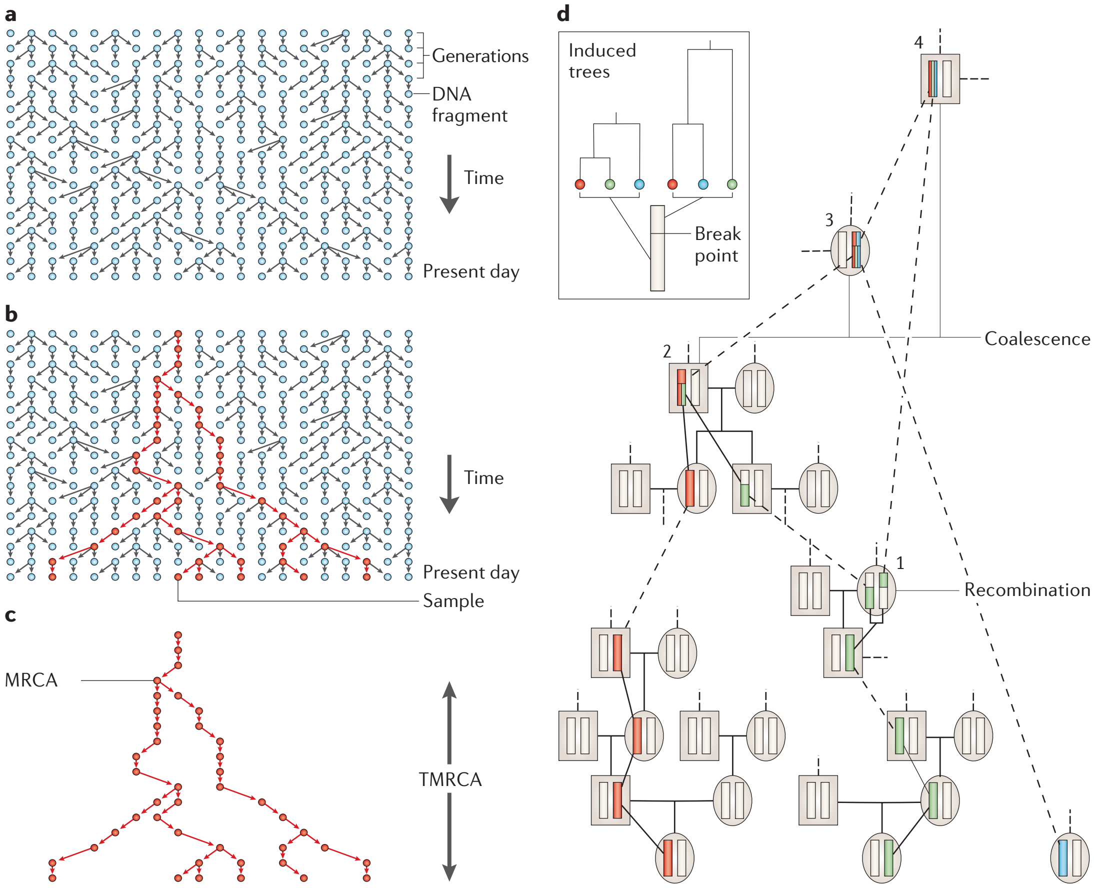
Panel **b** indicates the ancestry of a sample from the present day. In this example, six fragments, indicated in red, are sampled from the current generation. The ancestry of this sample is then traced back in time (that is, up the page), and is indicated in red.

Panel **c** highlights one of the key features of the coalescent: all information outside the ancestry of the sample of interest can be ignored. The coalescent provides a mathematical description of the ancestry of the sample. As we move back in time, the number of lines of ancestry decreases until, ultimately, a single line remains. The most recent fragment from which the entire sample is descended is known as the ‘most recent common ancestor’ (MRCA), whereas the time at which the MRCA appears is known as the ‘time to the most recent common ancestor’ (TMRCA).

Coalescent with recombination

The coalescent with recombination is illustrated in panel **d**. In such settings, lines bifurcate, as well as coalesce (join), as we move back in time. Here we show the genealogy for three copies of a fragment. By tracing the lineages back in time, we observe the following events: in event 1 the green lineage undergoes recombination and splits into two lineages, which are then traced separately; in event 2 one of the resulting green lineages coalesces with the red lineage, creating a segment that is partially ancestral to both green and red, and partially ancestral to red only; in event 3 the blue lineage coalesces with the lineage created by event 2, creating a segment that is partially ancestral to blue and red, and partially ancestral to all three colours; in event 4 the other part of the green lineage coalesces with the lineage created by event 3, creating a segment that is ancestral to all three colours in its entirety. As the inset shows, the recombination event induces different genealogical trees on either side of the break.

Coalescent methods have been reviewed extensively[20–22], and there are now book-length treatments[97,98] to which the reader is referred for further details.



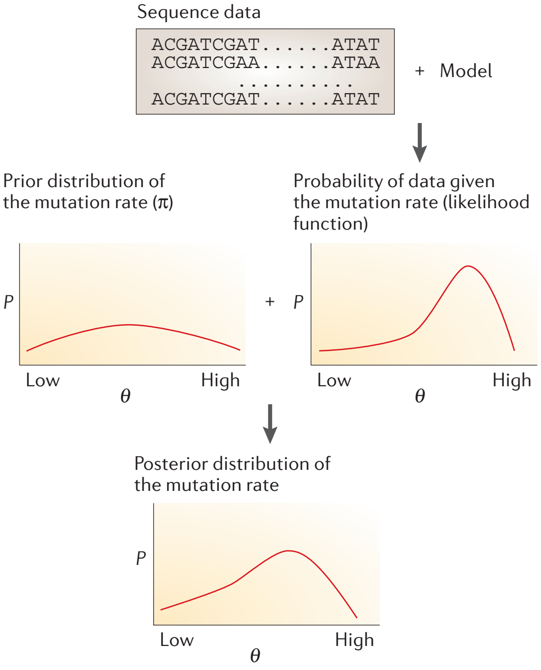
Panel **d** is modified with permission from REF. 89 © (2002) Elsevier.

Box 3 | Principles of model-based analysis

We demonstrate the principles of a model-based analysis using the example of estimating a mutation rate on the basis of a set of mitochondrial DNA (mtDNA) sequence data. As is common, the analysis is performed here in a Bayesian framework. The aim is to estimate the posterior distribution of a parameter, θ, in this case the DNA mutation rate, for a data set D.

For this example, the coalescent will typically be a reasonable choice of model. Prior information regarding the parameters of interest is incorporated into the prior distribution $\pi(\theta)$. We then calculate the posterior distribution for the parameter θ that is proportional to the product of the prior distribution and likelihood, that is, $f(\theta | D) \propto f(D|\theta)\pi(\theta)$, using one of the methods discussed in this article.

This calculation is shown in the figure. The three components are the data D (illustrated here by sequence data for some region), the coalescent model and the prior distribution for θ. The model is used to calculate the likelihood, that is, the probability (P) of the data (the y-axis of the graphs) over the range of possible mutation rates (θ; the x-axis of the graphs). This is then combined with the prior distribution to calculate the posterior distribution for the mutation rate.


Sequence data
ACGATCGAT......ATAT
ACGATCGAA......ATAA
....
ACGATCGAT......ATAT
+ Model
Prior distribution of the mutation rate (π)
Probability of data given the mutation rate (likelihood function)
P
Low
High
θ
+
P
Low
High
θ
Posterior distribution of the mutation rate
P
Low
High
θ

use results that have been obtained by simulation (or analysis) of this model using varying parameter values, combined with the properties of an observed data set, to estimate parameters. We approach this from a Bayesian perspective. A traditional alternative is to estimate parameters using the maximum-likelihood method. In a Bayesian framework, our prior knowledge of the parameters of interest is expressed in terms of a probability distribution known as the prior distribution. This is modified by the data to produce the posterior distribution, which summarizes our updated knowledge about the parameters conditional on the observed data. We give a formal statement of the model-based approach in BOX 3.

In the context of mtEve, we start with a set of mitochondrial DNA sequences obtained from a random sample of present-day individuals. We start with a model, in this case a coalescent with no recombination, and a prior distribution for population parameters, such as the mutation rate in the sequenced region. We then calculate the posterior distribution of the population parameters, the coalescent tree topology and the times of events on that topology. In this case, interest focuses on the posterior distribution of the TMRCA (that is, the mtEve of the sample[33–35]) and the mutation rate.

Most models of sequence evolution are sufficiently complicated that explicit calculation of the posterior distribution is impossible. In these cases, posterior distributions are usually obtained by using stochastic simulation methods. Put briefly, these methods involve repeatedly simulating the data under a range of parameter values, and then assessing how often the data are produced under the differing values of the parameter. We give a more detailed explanation in the following section.

Stochastic computation methods

Many approaches are available for constructing a posterior distribution. The choice of the most appropriate algorithm is determined by factors such as the complexity of the model and the size of the data set being considered. We now outline several of these common approaches. We also give some general guidelines regarding the limitations of the methods, and the conditions under which each might be an appropriate choice for a given data set.

Rejection algorithms. We begin with the simplest of the methods: rejection algorithms. This approach uses repeated simulation of data under plausible evolutionary scenarios. In layman's terms, a rejection algorithm repeatedly simulates data sets (D') using values of the parameter that are randomly sampled from the prior distribution. For each D' that is identical to the observed data D, the generating parameter values are stored (that is, that realization is 'accepted') and used to construct a posterior distribution for the parameters.

The main advantage of rejection methods is that, for most complicated genetics settings, it is far easier to simulate than to calculate. Many realistic models of evolution lead to distributions for which direct calculation is impossible, but which, given the recent improvements in computational power, can be simulated relatively easily. This leads to the easy development of rejection algorithms, with realistic evolutionary models, for the purposes of inference. An example is given in BOX 4.

Rejection algorithms such as those outlined above are known to perform poorly when the prior and posterior

Prior distribution
The distribution of likely parameter values before any data are examined.

Posterior distribution
The distribution that is proportional to the product of the likelihood and prior distribution.

distributions are markedly different, in particular when the degree of overlap in their coverage is small. Intuitively speaking, we spend a lot of time generating parameter values from the prior distribution, only to discover that they rarely lead to data that are anything like the observed *D*. For this reason, we pay close attention to the 'acceptance rate' of such algorithms: if the acceptance rate is so low that it takes an unreasonable amount of time to collect a large set of accepted parameter values, we use alternative methods.

As the quantity and, more importantly, complexity of biological data grows, any particular data set is, by nature of its complexity, unlikely. If we repeatedly simulate from a model under identical conditions the outcomes would be different to some degree. So, the probability of simulating *D′* data that are equal to *D* is very small. This has led to the adoption of more approximate methods. For example, we use summary statistics, as discussed in BOX 4 and the later section on approximate Bayesian computation.

It is worth noting that rejection algorithms can also be used for classical maximum-likelihood estimation, as opposed to the Bayesian approach that is described in BOX 4. One way to do this is to simulate observations with a uniform, or 'uninformative', prior distribution; because the likelihood is proportional to the posterior distribution in this case, the mode of the posterior distribution gives the maximum-likelihood estimator. An alternative is to use repeated simulation of data for a range of parameter values to approximate the likelihood[36]. Yet another approach is to use importance sampling, which is described in the next section.

Box 4 | **Rejection algorithms**

Basic features

Rejection methods use repeated simulation of the data as a method of inference. Loosely speaking, data are simulated under a range of values of the parameter η. At each step, if the data that are produced match the observed data, *D*, the parameter value that is being generated is 'accepted'. The set of accepted parameter values is then used to approximate the posterior distribution.

Details

A standard rejection algorithm would involve carrying out the following sequence of iterative steps:

Step 1. Sample the parameter η randomly from its prior distribution.

Step 2. Simulate data *D′* using the model with parameter η.

Step 3. Accept η if simulated data $D' = D$. Return to step 1.

The set of accepted η values is a random sample from the required posterior distribution[47].

Example application

The application of a rejection algorithm is illustrated here using the data of Ward *et al.*[6] Mitochondrial sequence data were collected for a sample of 63 members of the Nuu Chah Nulth tribe. The data consisted of 360 bp from hypervariable region I of the mitochondrial control region.

Nucleotide position in the control region

	69	88	91	106	124	149	162	166	190	194	200	209	233	247	251	255	267	271	275	296	301	302	304	319	339	344
ID:	T	C	C	G	C	T	C	T	G	T	C	C	C	C	G	C	C	C	T	G	T	T	C	T	T	A
1	.	.	.	.	.	.	.	C	A	.	T	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.
2	.	.	.	.	.	.	.	.	A	.	T	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	.	.	C	.
5	.	T	.	A	.	.	T	.	.	.	T	.	.	.	.	.	T	.	.	A	.	.	.	.	C	.
6	.	T	.	A	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	A	.	.	.	.	C	.
7	C	T	.	A	.	.	.	.	.	.	T	.	.	T	.	.	.	.	.	A	.	.	.	.	C	.
8	.	T	.	A	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	A	.	.	.	.	C	.
9	C	T	.	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	A	.	.	.	.	C	.
10	.	T	.	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	A	.	.	.	.	C	G
11	.	T	.	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	A	.	.	.	.	C	.
12	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	A	.	.	.	.	C	.
13	.	T	.	.	.	.	.	.	A	.	.	.	.	.	.	.	T	.	.	A	.	.	.	.	C	.
14	.	T	.	.	.	.	.	.	.	.	T	.	.	.	.	T	T	.	.	A	.	.	.	.	C	.
15	.	T	.	.	.	.	.	.	.	.	T	.	.	.	.	T	T	.	.	A	C	.	.	.	C	.
16	.	.	.	.	.	.	.	.	.	.	T	T	.	.	.	.	.	.	.	.	.	.	T	.	C	.
17	.	.	.	.	T	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	C	.	.	.	C	.
18	.	.	.	.	T	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	C	.	.	C	.
19	.	.	T	.	.	.	.	.	.	.	T	.	.	.	.	.	.	T	.	.	.	C	.	.	C	.
20	.	.	.	.	.	C	.	.	.	.	T	.	.	.	A	.	.	.	.	.	.	C	.	.	C	.
21	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	C	.	.	C	.
22	C	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.
23	.	.	.	.	.	.	.	.	.	.	T	T	.	.	.	.	.	.	C	.	.	C	.	.	.	.
24	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	C	.	.	C	T	.	.	.
25	.	.	.	.	.	.	.	.	.	.	T	T	.	.	.	.	.	.	C	.	.	C	T	C	.	.
26	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	C	.	.	C	T	C	.	.
27	.	.	.	.	.	.	.	C	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
28	.	.	.	.	.	.	.	C	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.

As shown in the figure, there were 28 distinct sequences observed, and 26 base positions showed variation within the sample. Dots indicate sequence identity with respect to the sequence shown at the top.

We demonstrate the use of rejection algorithms using the posterior distribution of the mutation rate, θ, and the time to the most recent common ancestor (TMRCA; see BOX 2), τ, of the data of Ward *et al.*[6] We let $\eta = (\theta,\tau)$ denote both the mutation rate for each 360 bp region and the TMRCA. For simplicity, rather than applying the rejection algorithm to the entire data set *D*, we use the number of segregating sites, κ, as a summary statistic of the data.

We proceed according to an iterative scheme.

Step 1. Sample the mutation rate θ from its prior distribution.

Step 2. Simulate a coalescent tree and superimpose mutations according to an appropriate mutation model. Count the number of segregating sites, κ', and record the height, τ, of the tree.

Step 3. Accept θ and τ if $\kappa' = \kappa$.

In this example we use a prior distribution for θ that is uniform over the range 0 to 100, and a standard coalescent model. The median of the prior distribution of the TMRCA is 1.71 and the median mutation rate is 50. The resulting posterior distribution for θ has a median of 7.2, whereas the posterior distribution for the TMRCA has a median of 1.55 on the coalescent timescale[70]. This shows a marked change in the mutation rate that is supported by the data, whereas the estimate of the TMRCA is only slightly reduced. We contrast these results with an analysis using the full data in BOX 6.

There are a number of variations of the rejection algorithm that is presented here; for examples, see REFS 32,70.

Figure modified with permission from REF. 6 © (1991) National Academy of Sciences.

Coverage
The range of values for which the probability is non-zero.

Summary statistics
A statistic that tries to capture a complicated data set in a simpler way. An example is the use of the number of segregating sites as a surrogate for a set of DNA fragments.

Importance sampling. Importance sampling is best used when we have some *a priori* idea of the nature of the posterior distribution. For example, one might have an idea of the range of reasonable parameter values. We then exploit this knowledge, which is framed in terms of an 'importance distribution', to improve computational efficiency.

In a Bayesian setting, importance sampling samples parameter values, which we denote by η, from an importance distribution $\xi(\eta)$, rather than the prior distribution $\pi(\eta)$. The simplest form of importance sampling proceeds in a similar way to rejection algorithms: each parameter value that is sampled from the importance distribution is accepted or rejected as before. However, because we are generating η from $\xi(\eta)$ rather than $\pi(\eta)$, we weight the accepted η values to compensate. Whereas in standard rejection methods each accepted η contributes a mass of weight 1 to the posterior distribution, it now contributes a mass proportional to $\pi(\eta)/\xi(\eta)$. For example, if $\xi(\eta) > \pi(\eta)$, we are sampling η more often than we would using the prior distribution, and we therefore down-weight the mass we give to each accepted use of η. If the importance-sampling distribution is well chosen (that is, sufficiently close to the (unknown) posterior), this strategy leads to a reduction in the variance between the estimated and actual summary statistics of the posterior distribution[37]. An example is shown in BOX 5.

Importance sampling is also used to evaluate the likelihood of the data, as a step towards calculating maximum-likelihood estimators. Such algorithms gained popularity in the population genetics field for estimating mutation and recombination rates. Griffiths and Tavaré developed importance-sampling algorithms that sampled from the collection of coalescent trees that might lead to a given data set[38–42]. These methods have been steadily improved by ensuring that the sampling is carried out according to an importance-sampling scheme that preferentially samples trees that are more likely to have resulted in the data[43–45], and have been generalized to a variety of other applications[46–51]. The Genetree software of Griffiths *et al.* (see online links) is widely used in this context.

If the importance-sampling distribution is well chosen, the algorithm will perform well, otherwise, it will perform poorly. Unfortunately, unless we have a good idea of the correct answer from some alternative source, it is not obvious whether the algorithm is working well. Once again there is significant scope for intuition when choosing the importance-sampling distribution[48]. The method is as much art as science.

Markov chain Monte Carlo methods. Another approach for constructing a posterior distribution, which is available when the explicit calculation of likelihoods is possible, is the Markov chain Monte Carlo (MCMC) approach. This method generates samples from the posterior distribution, but has the ability to learn from previous successes in the sense that, once a well-supported posterior region for the parameter is found, the algorithm, being Markovian, performs a more thorough exploration of that area. Therefore, MCMC algorithms are likely to perform better than rejection methods when the prior and posterior distributions are different. We discuss the methodology behind one type of MCMC — the Metropolis–Hastings[52,53] MCMC algorithm — in BOX 6.

Box 5 | **Importance sampling**

In brief, an importance sampling scheme is one in which the parameter values are sampled according to an importance distribution, rather than directly from the prior distribution. This importance distribution is chosen so as to make sampling more common at likely parameter values.

A simple example of the application of importance sampling is shown by using the same data set (the Nuu Chah Nulth data[6]) that was introduced in BOX 4. The results that have been obtained by two approaches (rejection algorithms and importance sampling) are compared. As in BOX 4, the aim is to construct a posterior distribution for the mutation rate θ and for the time to the most recent common ancestor (TMRCA) for a set of DNA sequences.

We begin by using a simple rejection method to create a benchmark to which we compare results. In this case, we assume a prior distribution for the total mutation rate θ across the 360 bp sequence. This is uniform on the interval [0,100] — in other words, all values of θ that lie within that range are assumed, *a priori*, to be equally likely; all values of θ that lie outside that range are assumed to have a probability of 0. On average, one θ value is accepted for every 34 randomly sampled values from the prior distribution. The results for θ and the TMRCA agree with those in BOX 4; for example, the median of the posterior distribution for the mutation rate is 7.2 for the entire region.

To demonstrate importance sampling, we now consider an analysis in which we use a prior distribution that is uniform in the range [18,28]. Using a rejection method, acceptances become rare, averaging one acceptance every 32,000 iterations. This is because the observed number of segregating sites is extremely unlikely for mutation rates in this range (the mutation rate that was found in BOX 4 and by the benchmark assay above was 7.2). To mitigate this problem, we use an importance-sampling scheme in which we sample values of θ according to an exponential distribution, with values ranging from 18 upwards and with a mean of 20. Acceptances now become more common, averaging one acceptance every 10,500 iterations, and, as must be the case, the results agree with those that were obtained using the prior distribution that is uniform over the range [18,28].

This simple example demonstrates the general feature that importance sampling can be used to improve the performance (in terms of acceptance rate in this example) in a context in which rejection methods perform poorly. In general, identifying a useful importance-sampling distribution is difficult. The weights of accepted observations can be used to assess the adequacy of the proposal distribution. Ideally, we do not want too many rejections and the variance of the weights should be low. For an extended discussion, see REF. 48.

Markov process
One in which the probability of the next state depends solely on the previous state, and not on the sequence of states before it.

Box 6 | An example of a Markov chain Monte Carlo method: the Metropolis–Hastings algorithm

Markov chain Monte Carlo (MCMC) methods generate observations from a posterior distribution by constructing a Markov chain with a stationary distribution that is the required posterior distribution. Simulation of the Markov chain results in observations that eventually have the correct distribution. We demonstrate this method using the Metropolis–Hastings algorithm, which is one of the simplest MCMC schemes. We use the mitochondrial sequence data of Ward *et al.*[6] (BOX 4) as an example. Once again, we aim to estimate the time to the most recent common ancestor (TMRCA) and the mutation rate.

The algorithm proceeds through a large number of iterations. At each iteration, the current configuration will consist of values for the parameters of interest (which, in this example, are the mutation rate and the coalescent tree topology) and a set of times of events on that topology. This time of events information is stored to help improve efficiency. At each iteration of the algorithm we propose a new set of parameter values, η'. In this example, we use η (or η') to denote both the mutation rate and the current tree, and the proposed new state will consist of a change to the tree and/or a change to the mutation rate. We then accept this new state (that is, the mutation rate and tree) with a probability h, known as the Hastings ratio, and defined as:

$$h = \min\left\{1, \frac{\mathrm{P}(D|\eta')\pi(\eta')q(\eta'\to\eta)}{\mathrm{P}(D|\eta)\pi(\eta)q(\eta\to\eta')}\right\} \qquad (1)$$

where $q(\eta \to \eta')$ denotes the probability of proposing a new state η' from the current state η; π is the prior distribution; D is the data set; P is the probability distribution; min is the minimum. If the new state is not accepted, the chain remains in the current state. The key to the efficient use of the MCMC scheme lies in the choice of the 'proposal kernel', q. If large changes are proposed, the data will typically be much less likely under the new state than under the existing state, and the proposed move will seldom be accepted (that is, the denominator will be greater than the numerator, and so the probability h of accepting a new state will be much less than 1). Therefore, changes are typically small, particularly with respect to the tree topology, in which one or two nodes of the tree are reconnected rather than changing the entire topology. Examples of how to do this can be found in REFS 60,79.

Subject to some conditions that ensure correct behaviour[99], once the algorithm has reached stationarity (and this is a key point), samples from the chain of η values represent draws from the required posterior distribution, $f(\eta|D)$.

In our mitochondrial Eve example, described in BOXES 4,5, we construct the posterior distribution for the TMRCA of the sample, and the mutation rate, using the heights of the coalescent tree and the mutation rate at each iteration. The median of the posterior distribution of the TMRCA is 0.62. The median for the mutation rate is 14.4 for the entire mitochondrial DNA region[61]. Note the contrast with the results in BOX 4, in which the median of the posterior distribution of the TMRCA was 1.55 and the median for the mutation rate was 7.2. Here we are using the full data, and so obtain the exact posterior distribution; by contrast, in BOX 4 we were using an approach based on summary statistics. The difference in the results is attributable to the loss of information that arises from summarizing the data. We discuss this more fully in the section on approximate Bayesian computation in the main text.

Although these algorithms have the advantage of producing samples from the posterior distribution, and are therefore widely used, several issues make their use difficult. First, it is difficult to assess whether the chain has reached stationarity. Theoretical work[54,55] has led to the introduction of several standard tool box diagnostics for this purpose, incorporated, for example, in the CODA package of Plummer *et al.* (see online links). Second, in direct contrast to rejection methods, consecutive parameter values are likely to be highly correlated; to overcome this limitation, the user will typically resort to sampling more widely spaced observations. This solution is not completely satisfactory because it is computationally inefficient. Third, in many applications it can be time-consuming to code and test such an algorithm.

The primary difficulty with MCMC algorithms, however, is the issue of mixing — that is, ensuring that the algorithm does not get 'stuck' in local maxima. Various solutions have been developed to deal with this problem. One of the simplest involves running several copies of the MCMC algorithm in parallel and starting from different points, with pairs of copies switching states from time-to-time[56]. Allowing copies to swap places occasionally means that the parameter space can be explored more efficiently. Other schemes involve augmenting the 'state–space' of the process: we add another variable to the space of parameters in such a way that it is easier for the algorithm to accept new states. For example, a useful idea is to add a 'temperature' to the process. In practice, this might involve mixing a 'hot' chain, which takes more frequent jumps, and a 'cool' chain, in which jumps are rarer. The addition of temperature allows the process to explore the parameter space with less risk of getting stuck; however, this greater efficiency occurs at the cost of the requirement for a more complicated algorithm. In some settings, a single process is run; in others, multiple parallel chains are used[48,57]. Owing to the additional complexity involved, these schemes have yet to be widely embraced within the genetics community.

Despite these caveats, MCMC algorithms are powerful and popular. In population genetics, a useful implementation is the LAMARC package of Kuhner *et al.* (see online links), which uses MCMC for maximum-likelihood estimation of evolutionary parameters, in various contexts, packaged in a user-friendly suite of programs. There are also many other purpose-built applications[58–66].

In the context of the mtEve example, an MCMC scheme is appropriate for analysing data when the number of observed SNPs, for instance, is relatively small (allowing calculations to occur in reasonable time) and when we are willing to assume a reasonably simple

Stationarity
The state in which a process has become independent of its starting position and has settled into its long-term behaviour. In an MCMC context, the process is typically assumed to be stationary at the end of a 'burn-in' period.

Local maxima
A local region in which a distribution takes a value that is higher than those taken at other nearby points, but which is lower than at least one value taken in some other, more distant region.

Box 7 | Approximate Bayesian computation methods

Approximate Bayesian computation (ABC) methods are motivated by a growing need to use more approximate models, or relatively simple summaries of full data sets, in order to keep the analysis tractable. They exist in a variety of forms, but here we focus on examples in which summaries of the data are used.

An example application

We return to the problem that is discussed in BOX 4. There we summarized the genetic variation in a sample of mitochondrial DNA sequences (the data) using the number of segregating sites as our summary statistic S. We saw that, using a rejection algorithm, the estimated time to the most recent common ancestor (TMRCA) had a median of 1.55, substantially different from the value of 0.62 that was obtained using the exact Markov chain Monte Carlo (MCMC) approach in BOX 6. Is this a consequence of the choice of summary statistic? To answer this, suppose that we summarize the data using both the number of segregating sites and the number of haplotypes. A rejection algorithm that tries to match both statistics has an acceptance rate of zero!

To overcome this limitation, we can relax the need for an exact match between the simulated and observed data using an ABC approach. Instead of trying to match both summary statistics, we could accept any iteration in which both statistics are within 2 of their values in the observed data. This leads to an acceptance rate of 1 in 10,000; the median estimate of TMRCA is now 0.64, close to the true answer of 0.62 that was obtained from the MCMC method given in BOX 6. To see whether this behaviour is representative, we need to consider an analysis in which an exact match is required. One approach is to use MCMC without likelihoods.

MCMC without likelihoods

In these methods, the step in BOX 6 that involved calculating the Hastings ratio, h, is replaced by two steps. In the first of these we simulate data. If the simulated and observed data are not identical we reject the current proposal. If the simulated data does match the observed data we proceed to the second step, which involves calculation of a simpler version of h (REF. 70). As with traditional MCMC, consecutive samples are correlated, so the caveats that apply to that method also apply here.

Applying this algorithm[70] results in a posterior median estimate for TMRCA of 0.55. Although this is closer to the truth than the answer that is obtained when using just S (TMRCA = 1.55), it is farther from the truth than when using the exact MCMC approach (TMRCA = 0.62) (BOX 5). This exemplifies the unintended effects that are possible when using summaries of the data.

mutation model (allowing calculation to be possible at all). Examples of such simple mutation models are those in which base pairs are assumed to mutate independently according to relatively simple mutation models. These might be, for example, that the mutant state is independent of the current state, or the new state at a base depends only on the current base at that position[67,68].

Approximate Bayesian computation

Each of the methods discussed so far can be computationally intensive. For example, rejection methods often fail because the acceptance rate is too low; this happens because (as explained above) it is difficult to simulate the observed data. In MCMC methods, the difficulty lies in evaluating the likelihood in a reasonable time. Considerations such as these motivate the use of more approximate methods. The approximation can occur in two areas. First, we no longer require an exact match between the observed and simulated data. Second, the underlying model can be simplified, but retain its key features.

First approximation: removing the need for an exact match between the simulated and observed data. In rejection methods, instead of requiring an exact match between the simulated and observed data, we accept the parameter values that correspond to any simulated data set that is sufficiently close to the observed data. Performance is now heavily dependent on the stringency of the required match between the simulated and observed data. An early example of this approach used in a biological context involved the inference of demographic parameters using microsatellite data on human Y chromosomes[69].

The comparison of simulated and observed data is often carried out using a set of summary statistics. An example is provided in BOX 4 in which the mitochondrial sequence data D were summarized by the number of segregating sites — an extremely simple summary. If the summary statistic S is sufficient for the parameter η then the posterior distribution of η given D is the same as its posterior distribution given S. Typically, S is of lower dimension than D, which makes the simulation methods much faster.

In complex problems, a low-dimensional sufficient statistic for the parameter of interest is usually unknown. This represents perhaps the main stumbling block in implementing summary methods, and there is a pressing need for new theory. If S is not sufficient for η, the resulting posterior is an approximation of the true posterior, and the closeness of the approximation is, *a priori*, unknown. The effects of summarizing the data can be hard to predict. Note the disparate estimates of TMRCA in BOXES 4,6: the MCMC method in BOX 6 used the full sequence data, whereas the rejection method in BOX 4, which summarized the data, produced a less accurate estimate. In the absence of a sufficient statistic, we rely on intuition to choose S, and then, perhaps, calibrate answers for a simpler form of the model from which we can find the exact posterior distribution[69,70].

In comparing summary statistics we might only accept iterations with exact matches between observed and simulated data. One alternative is to accept iterations with summary statistics that are sufficiently close to the target (that is, the observed data), which increases the acceptance rate. Another is to use every iteration, and post-process the output using a weighted linear regression[71]. The weight of each η is related to the distance between the data that are generated in that iteration and the observed data. This method can improve the properties of posterior estimates[71]. These approximate methods have become popularly known as the approximate Bayesian computation (ABC)[71]. An example is given in BOX 7.

These ideas can be exploited to construct MCMC algorithms when likelihoods cannot be calculated[70]. This approach is an appropriate choice when the data set is sufficiently large (and the mutation model is sufficiently complex) that explicit computation is slow or impossible, and when the posterior and prior distributions might be different. Although these 'no-likelihood' MCMC methods are new, and yet to be widely applied, they allow us to combine the ability of rejection methods to deal with intractable distributions with that of MCMC methods to explore local areas of high posterior probability with greater efficiency. Naturally, we also

Sufficiency

The statistic S is sufficient for the parameter η if the probability of the data, given S and η, does not depend on η.

Box 8 | **Summary of model-based analysis methods**

Properties or requirements of each method	Rejection	Importance sampling	Markov chain Monte Carlo	No-likelihood
Need to calculate likelihood	No	Maybe	Yes	No
Uses complex mutation models	Yes	Maybe	No	Yes
Independent samples	Yes	Maybe	No	No
‘Exact’ answer	Yes	Yes	Yes	No (typically)
‘Burn-in’ period	No	No	Yes	Yes
Ability to learn	No	No	Yes	Yes

The table shows a summary of the requirements and properties of each method we have discussed. Specifically, we show whether the method requires one to be able to calculate explicitly the probability of the data given the current configuration of the process; if this is the case, then the use of the method will be restricted to cases in which simple and potentially unrealistic models of mutation can reasonably be used.

We also indicate whether consecutive iterations of the process have the property of being independent, or whether there is correlation between such outputs. In the second case, one typically subsamples from consecutive outputs in an attempt to recover independence.

In this context, we note that although rejection and no-likelihood methods produce independent outputs, one might wait a long time for the next such output (as not every proposed new state is accepted; see BOXES 4,6). We also indicate whether the method gives exact samples from the required posterior distribution or whether it results in approximations to the same, and whether one has to wait until a ‘burn-in’ period has expired before sampling from the algorithm.

Finally, we indicate whether the algorithm has the ability to learn from the potential parameter values it has already explored, or whether it continues to sample from the same distribution throughout the course of the algorithm. For importance sampling, the answer to three of these questions depends on the particular form of algorithm that is used.

combine the disadvantages, in that we are seldom sure how close the estimated posterior is to the true posterior, and we no longer obtain independent draws from that posterior. Furthermore, preliminary experience with these methods suggests that they can have poor mixing properties.

There is, most definitely, no such thing as a free lunch in this field. Further work is now beginning to emerge[72]. ABC schemes allow the use of more complex and realistic evolutionary models.

Second approximation: simplifying the models. An example of approximating by simplifying the model occurs when simulating coalescent data with recombination. When simulating haplotype data over relatively short regions (of the order of 100 kb, for instance) it has been traditional to use the coalescent to simulate the ancestry of the sample. As recombination rates increase, lines of ancestry split many times until the size of the graph prevents it from being stored in computer memory. McVean and Cardin[24] introduced an approximation to the coalescent that efficiently simulates data in a genome-wide context. Their method builds on a clever ‘along the chromosome’ construction of the coalescent, attributable to Wiuf and Hein[73,74], in which the full coalescent graph is constructed as a set of simple coalescent trees as we move from one end of the chromosome to the other. This approach avoids some of the complexities that are inherent in the construction of the full graph, and has itself been refined and implemented as distributable code by Marjoram and Wall[25].

Haplotype
The sequence of bases along a single copy of (typically, part of) a chromosome.

Discussion and further perspectives

A wide variety of computational methods have been developed for the analysis of genetic data. We have focused on model-based approaches and in BOX 8 we show a summary of the properties of those we have discussed. We conclude with some comments on where these approaches are headed.

What are models for? As George Box noted: “All models are wrong, some are useful.”[75] The main utility of models in population genetics is to support an intuition about the influence of different forces on the structure of the genetic variation that is observed in the population. Models also provide a way of assessing the properties of estimators of parameters of interest. For example, it was noted early on[27,28] that estimators of mutation parameters typically have a much larger variance than would be expected using ‘standard’ statistical theory. This reflects the dependence between observations that is attributable to the shared common ancestry of the sample. The fact that all individuals share ancestry with each other means that the properties of sample members are correlated, and this increases the variance of the estimators (compared with a set of independent observations of equal size). The coalescent can be used to quantify the extent of this correlation.

What will be the role of models in the future? In response to increases in the quantity and complexity of molecular data, more detailed biological models will be developed. Such models will explicitly describe the details of the molecular processes that produce the data. From the perspective of inference, precise formal analysis of such models is likely to be extremely difficult, and the focus is likely to change in two related ways: through the development of simpler models that capture the essential features (such as the effects of dependence) of the more complicated ones, and through the development of simpler methods of analysis, such as ABC.

A related issue is ‘goodness of fit’. Do the data produced by the model look like the observed data? This is a question that is seldom addressed clearly. It is likely that, as data become richer, the relatively simplistic models that have commonly been used to date will be shown to be inadequate. Some effort will be required to develop more complex models that remain tractable, or to find a combination of parameters that can accurately simulate data at this new, higher degree of detail. An example of this is found in REF. 76, in which the authors find parameter values for a coalescent model that reflect key features of observed variation in the human genome.

Conclusion

In this review we have given a necessarily selective overview of the computational methods that are used in population genetics. In particular, we have described only the simplest methods. In practice, different stochastic computational techniques are often combined to address a given problem. There are numerous generalizations of the methods we have described. Liu[48] and Robert and Casella[77] provide comprehensive coverage of the general area.

For any problem, it is generally the case that many of the methods could be applied. The devil is in the details, and it is those details that determine which method is the most appropriate choice. These methods are generally applied to more complex problems than those discussed here. Although there are several standard tools available to facilitate standard applications of these methods[78], the complexity of population genetics models means that it is rarely practical to use these tools. Combining the intuition that is provided by complex stochastic models with the judicious use of simulation methods for inference will dominate the field from now on.

1. Hubby, L. & Lewontin, R. C. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics* **54**, 577–594 (1966).
2. Jeffreys, A. J. DNA sequence variants in the Gγ-, Aγ-, Δ- and β-globin genes. *Cell* **18**, 1–10 (1979).
3. Kan, Y. W. & Dozy, A. M. Polymorphism of DNA sequence adjacent to human β-globin structural gene: relationship to sickle mutation. *Proc. Natl Acad. Sci. USA* **75**, 5631–5635 (1978).
4. Kreitman, M. Nucleotide polymorphism at the alcohol-dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304**, 412–417 (1983).
5. Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).
6. Ward, R. H., Frazier, B. L., Dew-Jager, K. & Pääbo, S. Extensive mitochondrial diversity within a single Amerindian tribe. *Proc. Natl Acad. Sci. USA* **88**, 8720–8724 (1991).
7. Whitfield, L. S., Sulston, J. E. & Goodfellow, P. N. Sequence variation of the human Y chromosome. *Nature* **378**, 379–380 (1995).
8. Dorit, R. L., Akashi, H. & Gilbert, W. Absence of polymorphism at the ZFY locus on the human Y chromosome. *Science* **268**, 1183–1185 (1995).
9. Jorde, L. B. *et al*. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y chromosome data. *Am. J. Hum. Genet.* **66**, 979–988 (2000).
10. Rosenberg, N. A. *et al*. Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
11. Nordborg, M. *et al*. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3**, 1289–1299 (2005).
12. Altshuler, D. *et al*. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
13. Yu, J. & Buckler, E. S. Genetic association mapping and genome organization of maize. *Curr. Opin. Biotechnol.* **17**, 155–160 (2006).
14. Provine, W. B. *The Origins of Theoretical Population Genetics* (Univ. Chicago Press, Chicago; London, 1971).
15. Ewens, W. J. *Mathematical Population Genetics* (Springer, Berlin; Heidelberg; New York, 1979).
Describes the state-of-the-art in population genetics theory before the appearance of the coalescent.
16. Slatkin, M. & Hudson, R. R. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**, 555–562 (1991).
17. Kingman, J. F. C. On the genealogy of large populations. *J. Appl. Prob.* **19A**, 27–43 (1982).
Introduces the coalescent as a way of exploiting ancestry in population genetics models.
18. Kingman, J. F. C. The coalescent. *Stochastic Proc. App.* **13**, 235–248 (1982).
19. Hudson, R. R. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**, 183–201 (1983).
Introduces the coalescent with recombination.
20. Hudson, R. R. in *Oxford Surveys in Evolutionary Biology* (eds Futuyma, D. & Antonovics, J.) (Oxford Univ. Press, New York, 1991).
21. Donnelly, P. & Tavaré, S. Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* **29**, 401–421 (1995).
22. Nordborg, M. in *Handbook of Statistical Genetics* (eds Balding, D. J., Bishop, M. J. & Cannings, C.) (John Wiley & Sons, New York, 2001).
23. Hudson, R. R. Generating samples under a Wright–Fisher neutral model. *Bioinformatics* **18**, 337–338 (2002).
24. McVean, G. A. T. & Cardin, N. J. Approximating the coalescent with recombination. *Philos. Trans. R. Soc. Lond. B* **360**, 1387–1393 (2005).
25. Marjoram, P. & Wall, J. D. Fast 'coalescent' simulation. *BMC Genetics* **7**, 16 (2006).
26. Peng, B. & Kimmel, M. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics* **21**, 3686–3687 (2005).
27. Ewens, W. J. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3**, 87–112 (1972).
The first rigorous statistical treatment of inference for molecular population genetics data.
28. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
A classic paper that introduces the number of segregating sites as the basis of an efficient estimator for mutation rate.
29. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
30. Griffiths, R. C. &. Tavaré, S. The age of a mutation in a general coalescent tree. *Stochastic Models* **14**, 273–295 (1998).
31. Slatkin, M. & Rannala, B. Estimating allele age. *Annu. Rev. Genomics Hum. Genet.* **1**, 225–249 (2000).
32. Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. Inferring coalescence times for molecular sequence data. *Genetics* **145**, 505–518 (1997).
33. Tang, H., Siegmund, D. O., Shen, P., Oefner, P. J. & Feldman, M. W. Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. *Genetics* **161**, 447–459 (2002).
34. Meligkotsidou, L. & Fearnhead, P. Maximum-likelihood estimation of coalescence times in genealogical trees. *Genetics* **171**, 2073–2084 (2005).
35. Tavaré, S. in *Case Studies in Mathematical Modeling: Ecology, Physiology, and Cell Biology* (eds Othmer, H. G. *et al*.) (Prentice–Hall, New Jersey,1997).
36. Diggle, P. J. & Gratton, R. J. Monte Carlo methods of inference for implicit statistical models. *J. R. Stat. Soc. B* **46**, 193–227 (1984).
37. Ripley, B. D. *Stochastic Simulation* (John Wiley & Sons, New York, 1987).
38. Griffiths, R. C. & Tavaré, S. Simulating probability distributions in the coalescent. *Theor. Popul. Biol.* **46**, 131–159 (1994).
39. Griffiths, R. C. & Tavaré, S. Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. Lond. B* **344**, 403–410 (1994).
40. Griffiths, R. C. & Tavaré, S. Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math. Biosci.* **127**, 77–98 (1995).
41. Griffiths, R. C. & Tavaré, S. Ancestral inference in population genetics. *Stat. Sci.* **9**, 307–319 (1994).
42. Griffiths, R. C. & Tavaré, S. Monte Carlo inference methods in population genetics. *Math. Comput. Model.* **23**, 141–158 (1996).
43. Felsenstein, J., Kuhner, M., Yamato, J. & Beerli, P. in *Statistics in Molecular Biology and Genetics* (ed. Seillier-Moiseiwitsch, F.) 163–185 (Hayward, California, 1999).
44. Stephens, M. & Donnelly, P. Inference in molecular population genetics. *J. R. Stat. Soc. B* **62**, 605–655 (2000).
45. De Iorio, M. & Griffiths, R. C. Importance sampling on coalescent histories. I. *Adv. Appl. Prob.* **36**, 417–433 (2004).
46. Griffiths, R. C. & Marjoram, P. Ancestral inference from samples of DNA sequences with recombination. *J. Comp. Biol.* **3**, 479–502 (1996).
47. Stephens, M. in *Handbook of Statistical Genetics* (eds Balding, D. J., Bishop, M. J. & Cannings, C.) 213–238 (John Wiley & Sons, New York, 2001).
48. Liu, J. S. *Monte Carlo Strategies in Scientific Computing* (Springer, New York, 2001).
49. De Iorio, M. & Griffiths, R. C. Importance sampling on coalescent histories. II. Subdivided population models. *Adv. Appl. Prob.* **36**, 434–454 (2004).
50. De Iorio, M., Griffiths, R. C., Lebois, R. & Rousset, F. Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theor. Popul. Biol.* **68**, 41–53 (2005).
51. Chen, Y. & Xie, J. Stopping-time resampling for sequential Monte Carlo methods. *J. R. Stat. Soc. B* **67**, 199–217 (2005).
52. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
53. Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109 (1970).
54. Cowles, M. K. & Carlin, B. P. Markov chain Monte Carlo diagnostics: a comparative review. *J. Am. Stat. Assoc.* **91**, 883–904 (1995).
55. Brooks, S. P. & Roberts, G. O. Assessing convergence of Markov chain Monte Carlo algorithms. *Stat. Comput.* **8**, 319–335 (1998).
56. Wilson, I. J. & Balding, D. J. Genealogical inference from microsatellite data. *Genetics* **150**, 499–510 (1998).
57. Nielsen, R. & Palsboll, P. J. Single-locus tests of microsatellite evolution: multi-step mutations and constraints on allele size. *Mol. Phylogenet. Evol.* **11**, 477–484 (1999).
58. Markovtsova, L., Marjoram, P. & Tavaré, S. The age of a unique event polymorphism. *Genetics* **156**, 401–409 (2000).
59. Markovtsova, L., Marjoram, P. & Tavaré, S. The effects of rate variation on ancestral inference in the coalescent. *Genetics* **156**, 1427–1436 (2000).
60. Nielsen, R. & Wakeley, J. W. Distinguishing migration from isolation: an MCMC approach. *Genetics* **158**, 885–896 (2001).
61. Fearnhead, P. & Donnelly, P. Estimating recombination rates from population genetic data. *Genetics* **159**, 1299–1318 (2001).
62. Fearnhead, P. & Donnelly, P. Approximate likelihood methods for estimating local recombination rates. *J. R. Stat. Soc. B* **64**, 657–680 (2002).
63. Li, N. & Stephens, M. Modelling linkage disequilibrium, and identifying recombination hotspots using SNP data. *Genetics* **165**, 2213–2233 (2003).
An early application of the ABC idea; it is used here to construct tractable approximations to more complex evolutionary models.
64. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
65. Thorne, J. L., Kishino, H. & Felsenstein, J. Inching towards reality: an improved likelihood model of sequence evolution. *J. Mol. Evol.* **34**, 3–16 (1992).
66. Felsenstein, J. Evolutionary trees from DNA sequence data: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).

67. Geyer, C. J. in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (ed. Keramidas, E. M.) (Interface Foundation, Fairfax Station, 1991).

68. Geyer, C. J. & Thompson, E. A. Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Am. Stat. Assoc.* **90**, 909–920 (1995).

69. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* **16**, 1791–1798 (1999).

70. Marjoram, P., Molitor, J., Plagnol, V. & Tavaré, S. Markov chain Monte Carlo without likelihoods. *Proc. Natl Acad. Sci. USA* **100**, 15324–15328 (2003).

71. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002). **Coins the term approximate Bayesian computation, and applies it to microsatellite data.**

72. Bortot, P., Coles, S. G. & Sisson, S. A. Inference for stereological extremes. *J. Am. Stat. Assoc.* (in the press).

73. Wiuf, C. & Hein, J. Recombination as a point process along sequences. *Theor. Popul. Biol.* **55**, 248–259 (1999).

74. Wiuf, C. & Hein, J. The ancestry of a sample of sequences subject to recombination. *Genetics* **151**, 1217–1228 (1999). **References 73 and 74 present an elegant construction of the coalescent in the presence of recombination.**

75. Box, G. E. P. in *Robustness in Statistics* (eds Launer, R. L. & Wilkinson, G. N.) (Academic Press, New York, 1979).

76. Schaffner, S. F. *et al*. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15**, 1576–1583 (2005). **A comprehensive study that shows that the coalescent is a good model for complex evolutionary data.**

77. Robert, C. P. & Casella, G. *Monte Carlo Statistical Methods* (Springer, New York, 2004).

78. Spiegelhalter, D. J., Thomas, A., Best, N. & Lunn, D. *WinBUGS Version 1.4 User Manual* [online], < http://www.mrc-bsu.cam.ac.uk/bugs > (2003).

79. Kuhner, M., Yamato, J. & Felsenstein, J. Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. *Genetics* **140**, 1421–1430 (1995).

80. Wall, J. D. A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.* **17**, 156–163 (2000).

81. Smith, N. G. C. & Fearnhead, P. A comparison of three estimators of the population-scaled recombination rate: accuracy and robustness. *Genetics* **171**, 2051–2062 (2005).

82. Hudson, R. R. Two-locus sampling distributions and their applications. *Genetics* **159**, 1805–1817 (2001).

83. McVean, G. A. T. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).

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

85. Kuhner, M., Yamato, J. & Felsenstein, J. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149**, 429–434 (1998).

86. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000). **Introduces a widely used method for inferring population structure.**

87. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).

88. Pollinger, J. P. *et al*. Selective sweep mapping of genes with large phenotypic effects. *Genome Res.* **15**, 1809–1819 (2006).

89. Nordborg, M. & Tavaré, S. Linkage disequilibrium: what history has to tell us. *Trends Genet.* **18**, 83–90 (2002).

90. Stephens, M., Smith, N. J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–989 (2001). **Describes an elegant use of the coalescent to infer haplotype phase from SNP data.**

91. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–644 (2006).

92. Crawford, D. C. *et al*. Evidence for substantial fine-scale variation in the recombination rate across the human genome. *Nature Genet.* **36**, 700–706 (2004).

93. Fearnhead, P. & Smith, N. G. C. A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes. *Am. J. Hum. Genet.* **77**, 781–794 (2005).

94. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321–324 (2005).

95. Tiemann-Boege, I., Calabrese, P., Cochran, D. M., Sokol, R. & Arnheim, N. High resolution recombination patterns in a region of human chromosome 21 measured by sperm typing. *PLoS Genet.* **2**, e70 (2006).

96. Balding, D. J. A tutorial on statistical methods for population association studies. *Nature Rev. Genet.* **7**, 781–791 (2006).

97. Hein, J., Schierup, M. H. & Wiuf, C. *Gene Genealogies, Variation and Evolution* (Oxford Univ. Press, New York, 2005).

98. Tavaré, S. in *Lectures on Probability Theory and Statistics. Ecole d'Etés de Probabilité de Saint-Flour XXXI — 2001* (ed. Picard, J.) (Springer, Berlin; Heidelberg, 2004).

99. Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. *Markov chain Monte Carlo in Practice* (Chapman & Hall, London, 1996).

Acknowledgements

The authors were supported in part by two grants from the US National Institutes of Health. S.T. is a Royal Society-Wolfson Research Merit Award holder. We thank the reviewers for helpful comments on an earlier version of the manuscript.

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

CODA: http://cran.r-project.org/src/contrib/Descriptions/coda.html
FPG: http://lifesci.rutgers.edu/~heylab/ProgramsandData/Programs/FPG/FPG_Documentation.htm
Genetree: http://www.stats.ox.ac.uk/mathgen/software.html
LAMARC v 2.0: http://evolution.gs.washington.edu/lamarc
SIMCOAL: http://cmpg.unibe.ch/software/simcoal
SimuPOP: http://simupop.sourceforge.net
The International HapMap Project: http://www.hapmap.org
Access to this links box is available online.