THE STUDY OF STRUCTURED POPULATIONS — NEW HOPE FOR A DIFFICULT AND DIVIDED SCIENCE

Jody Hey and Carlos A. Machado‡*

Natural populations, including those of humans, have complex geographies and histories. Studying how they evolve is difficult, but it is possible with population-based DNA sequence data. However, the study of structured populations is divided by two distinct schools of thought and analysis. The phylogeographic approach is fundamentally graphical and begins with a gene-tree estimate. By contrast, the more traditional approach of using summary statistics is fundamentally mathematical. Both approaches have limitations, but there is promise in newer probabilistic methods that offer the flexibility and data exploitation of the phylogeographic approach in an explicitly model-based mathematical framework.

DEMOGRAPHIC HISTORY
The reproductive history of a population or group of populations. This can include population sizes, sex ratios, migration rates, population-splitting events, variation in reproductive rates and times among organisms, as well as variation over time in all of these quantities.

**Department of Genetics, Rutgers the State University of New Jersey, 604 Allison Road, Piscataway, New Jersey 08854, USA.*
‡Department of Ecology and Evolutionary Biology, University of Arizona, 1041 East Lowell Street, BSW 308, Tucson, Arizona 85721, USA.
Correspondence to: J.H.
e-mail: hey@biology.rutgers.edu
doi:10.1038/nrg1112

In the early twentieth century, when the new science of genetics was picking up steam, the big question was whether Mendel's rules of inheritance could be reconciled with a Darwinian theory of evolution. In the course of finding out that they were compatible, the new science of population genetics was born[1]. Today, as at the beginning, population genetics is the study of how evolution works as a genetic process in natural populations. It is a difficult science, which is often highly mathematical in theory and approximate in application. Real populations are rarely simple, so it is difficult to research and develop theories about them. Natural populations are also dynamic in many dimensions: over time they change in size, density and location, and over space they can fragment into several populations and join with others. An awareness of those complexities, and a desire to have evolutionary models that are as realistic as possible, has led many population geneticists to focus their efforts on what has come to be called the 'structure' of natural populations. This field has grown with the availability of comparative DNA sequence data from natural populations. However, growth has proceeded in two different directions: the first relies on graphical depictions of the branching evolutionary history of DNA sequences, whereas the second relies on mathematical models of population structure.

This review does not cover all aspects of population structure; instead, we focus on models and methods for the study of population structure that use DNA sequence data, which we illustrate with examples from the literature on human populations. Our species has had a complex DEMOGRAPHIC HISTORY and provides examples of many kinds of population structure. Also, a considerable part of modern medical genetics relies on an understanding of human demographic history, so there is a strong demand for high-quality human population-genetic research.

Starting simply

To gain a starting purchase, we need a simple population model that does not have structure. Sewall Wright and Ronald A. Fisher independently described what has come to be the standard simple population model for most circumstances[2,3]. The main feature of the Wright–Fisher model (BOX 1) is the persistence of a single population of constant size, with random mating among individuals (panmixia). For most purposes it is assumed that the population has persisted for a long period (literally an infinite length of time, for mathematical purposes).

Box 1 | The Wright–Fisher model

Sewall Wright and Ronald A. Fisher were the pioneers of population genetics. Independently, they each made use of a simple mathematical representation of an idealized population[2,3]. A Wright–Fisher population has the following main components:

- A constant population size of *N* diploid individuals
- Non-overlapping generations, so that all individuals die following reproduction
- Random mating among individuals
- A random number of offspring per individual, which follows a POISSON DISTRIBUTION

By themselves, these assumptions are suitable for modelling the processes of genetic drift and gene-tree depths. By adding further components for particular problems, it is possible to use the model to study natural selection and population structure. For example, it is not difficult to include neutral mutations to model genetic variation in the absence of natural selection. A similar model with overlapping generations was developed by Moran[93].

Under these circumstances, the population can be represented by a single quantity — the population size (generally denoted by '*N*').

Real populations are different from Wright–Fisher populations, as they often have complex geographies and consist of many populations that are connected by GENE EXCHANGE. Individuals in most real populations are more likely to reproduce with nearby individuals than with distant individuals. FIG. 1 depicts four principal classes of models that have been developed to consider population structure. Analysis of these models has greatly helped our understanding of how geography and limited gene exchange can impact patterns of genetic variation. However, in most applications the models contain a partly hidden assumption that overlooks another important way in which real populations depart from a simple ideal. The assumption is that the populations have been evolving in the model for a long time and have reached a 'steady state' or 'equilibrium' pattern of variation in and between sub-populations. In general, the equilibrium arises as a balance between the actions of mutation and GENETIC DRIFT (which tend to make populations different) and gene exchange (which makes them more similar)[4]. Such models are useful for determining the types of pattern of variation in and between populations that are expected as a result of the structure model, assuming that the model has been in place for a long time.

A different class of non-equilibrium models have parameters that can change as a function of time[5–8]. For example, population sizes might change over time, as might the number of populations and the rates of gene exchange. There is generally not a steady-state pattern of genetic variation in non-equilibrium models because the details of the model are different at different points in time.

Equilibrium models are simpler than non-equilibrium models and seemed appropriate before DNA sequence data became commonplace. At that time, the basic genetic contrast between copies of a gene was based on the idea of 'alleles', in which two copies of a gene were either identical or different. However, when several DNA sequence copies of a gene are obtained from a natural population, each pair can differ in many ways. With allelic data, the basic measurements are allele frequencies and, in the case of diploid populations, departures from HARDY-WEINBERG expectations. These measurements lend themselves well to the fitting of equilibrium models, but generally do not have the information content that is needed to assess whether model parameters have changed over time. By contrast, a set of DNA sequences from one gene often shows a large number of polymorphic sites with high information content that will provide greater scope for the interpretation of changing population-model parameters.

A divided science — to tree or not to tree?

When population geneticists began collecting DNA sequence data from natural populations in the late 1970s, the field underwent a notable shift, and a split developed between two main schools of thought. Unlike allelic data, DNA sequence data can be readily applied to the calculation of EVOLUTIONARY TREES, and some investigators quickly realized that these kinds of trees could be a basis for the study of population history. Such GENE TREES are different from traditional PHYLOGENETIC TREES (the former describe the pattern of DNA ancestry in a population, the latter a pattern of taxon ancestry), but the tools that SYSTEMATICS uses for building evolutionary trees could be used to build population gene trees.

Using DNA sequence data to build gene trees was a different way of doing population genetics. Instead of focusing on numbers and mathematical models, this new tree-based framework was fundamentally graphical. The main pioneers in this field were Wesley Brown and John Avise, and the phylogenetic approach, as applied to problems of population structure, became known as phylogeography[9]. The growing field took advantage of advances in methods for the estimation of evolutionary trees and has become a well-recognized discipline in evolutionary biology[10,11].

But what became of traditional mathematical population genetics as the popularity of the phylogenetic approach grew? The field has thrived separately from phylogeography. The high information content in DNA sequences, and the historical information that is shown in patterns of DNA sequence variation, have opened up new areas of mathematical model development, which are commonly referred to as 'coalescent modelling' in reference to a family of mathematical models of the common ancestry of DNA sequences[12–15]. However, unlike tree-based methods, COALESCENT THEORY is generally directed at the effect of evolutionary forces on levels and patterns of DNA sequence variation. An important focus of these methods has been to develop ESTIMATORS of population-genetic parameters, such as the population mutation rate (generally denoted by θ; see later for further discussion), the time since two populations separated from each other and the rate of migration between populations.

The body of theory and statistical tools that constitute modern mathematical population genetics lacks a familiar all-encompassing name, such as 'phylogeography', so hereafter we identify these methods by their common reliance on SUMMARY STATISTICS. Unlike phylogeography, summary-statistic methods make no use of the genealogy that underlies a data set. For these methods, the gene

POISSON DISTRIBUTION
A probability distribution that is commonly used to describe the frequency at which similar but independent events can be expected to occur over a given period of time.

GENE EXCHANGE
The process by which genetic material is shared among organisms, which can occur through sexual reproduction or lateral genetic transfer.

GENETIC DRIFT
Random changes in gene frequency in a population that occur when a finite number of progeny are formed by the random sampling of gametes from the parents.

HARDY-WEINBERG
A classical mathematical principle in population genetics that describes the expected frequencies of genotypes for one locus after one generation of random mating if the allele frequencies in the parents are known.

EVOLUTIONARY TREE
A graph or branching diagram that describes the pattern of evolutionary ancestry (historical relationships) among a group of organisms.

GENE TREE
A graph or branching diagram that describes the pattern of ancestry among homologous DNA sequences from different individuals of a population or species.

PHYLOGENETIC TREE
A graph or branching diagram that describes the pattern of ancestry among different species or other taxa.

tree is neither important nor is it considered for parameter estimation. They begin, not with an evolutionary tree, but by summarizing some aspect of the data (for example, by counting the number of variable sites in or between populations). Whereas the phylogeographic approach does not rely on any explicit historical demographic model, summary statistics generally have little meaning unless they are considered together with the model under which they were calculated. FIG. 2 provides an example of the contrast between the starting points of the two approaches.

Gene-tree-based methods and applications

The underlying goal of tree-based phylogeographic analyses is to discover the history of related populations (from one or several closely related species) based on the depth and branching patterns of a gene tree of DNA sequences[9,10]. So far, most studies have focused on mitochondrial DNA (mtDNA)[11,16], which in most organisms is inherited from only one parent as a single non-recombining unit. The mtDNA is also often convenient because it has evolved quickly, which allows access to more recent population history. Increasingly, nuclear genes have also been included in phylogeographic studies[16–22], although the interpretation of such genealogies or haplotype trees is difficult if there has been recombination, as is common in nuclear genes.

Some classic mtDNA-based phylogeographic studies of humans are a good example of the benefits and drawbacks of the phylogeographic approach[23,24]. These studies generated excitement and controversy as they indicated that present populations descended from a relatively small population in Africa. The estimated common ancestor for the mtDNA genome was believed to have existed ~200,000 years ago, which is consistent with an effective population size of ~10,000 individuals. The two principal observations were that the tree depth, relative to a chimpanzee OUTGROUP, indicated a recent pattern of common ancestry, and that the most basal splits in the tree had only African descendants. Much of the controversy surrounding these claims arose from mtDNA sequence HOMOPLASY, which greatly hindered gene-tree estimation. As interpretations of recent African ancestry hinged on estimated geographical locations of ancestral mtDNAs, doubts about the gene-tree branching pattern lead to doubts about whether the data supported an African location of the human mtDNA ancestor[25–27]. Another concern stems from the fact that mtDNA is inherited effectively as a single LINKAGE BLOCK owing to the absence of recombination, so other genes could be expected to have different histories[28].

Phylogeographic analysis begins with the estimation of a gene tree, however it can be difficult to use a tree for statistical tests of specific hypotheses. For this purpose, many turn to a method that was developed by Alan Templeton and colleagues, which is called nested-clade analysis (NCA) (BOX 2). NCA is designed to distinguish different historical processes that might have influenced the geographic distribution of genetic variation as seen in a tree[29,30]. Such processes might include gene exchange, isolation by distance, past fragmentation events (such as ALLOPATRIC DIFFERENTIATION) and range expansions. Recently, Templeton applied this approach to questions about the spread of human populations out of Africa[31]. Data from several genes indicate that modern populations might be the result of a complex history that included several migrations out of Africa over the past several hundred thousand years — an interpretation that is in conflict with the idea that the migrations that gave rise to present populations outside of Africa did not begin until ~100,000 years ago[24,32].

There are two components to NCA: a statistical test of geographic structure, and a protocol for inference of the causes of that structure if it is found. The latter has

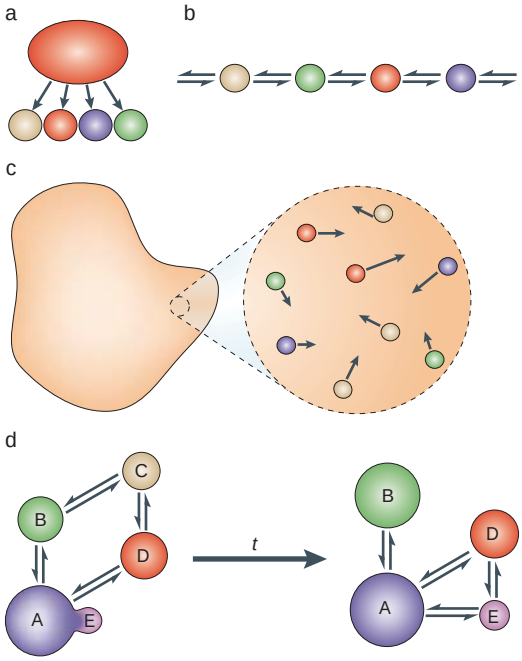


Figure 1 | **Models of population structure. a** | Island model of migration. The simplest island model has a mainland population with migration to one or more island populations. Island models can vary widely in numbers of populations, sizes of populations and rates of gene exchange. Island models are good for understanding the effects of small population size and limited gene flow on rates of genetic drift and levels of divergence between island populations[98]. **b** | Stepping-stone model. Unlike island models, this class of models specifically include a spatial element, with individual populations only able to exchange genes with adjacent populations. Stepping-stone models can be one-dimensional (with populations in a line, as shown), two- or three-dimensional[99]. **c** | Isolation by distance model[100,101]. If a stepping-stone model is taken to the extreme, then every individual is restricted in its local movement (or the distance that its genes can travel). This leads to the idea of a population that is evenly distributed over a landscape (two-dimensional in the figure, with each individual being capable of moving only a short distance, on average). If movement distances per generation are short, then on average, individuals are much more closely related to nearby individuals than to distant individuals. **d** | Metapopulation model. In nature, not only do individuals move between populations, but also individual populations come and go over time (*t*) with the founding and extinction of entire populations being an important component of population structure[102,103].

SYSTEMATICS
A branch of biology that deals with the classification of living organisms on the basis of their evolutionary relationships. This differs from 'taxonomy' as organisms are grouped on the basis of shared ancestry, not just on their similarities (which might or might not correspond to shared evolutionary history).

COALESCENT THEORY
A mathematical approach that models the depths of gene trees for samples that are drawn from one or more closely related populations.

ESTIMATOR
A method for calculating an estimate of a parameter in a model.

SUMMARY STATISTIC
A number that is calculated from a data set, which represents much of the information in the data. For a set of DNA sequences, one commonly used summary statistic is S, which represents the number of variable sites in the sample. Summary statistics are often easier to use to fit models to data than would be the case with the data itself.

OUTGROUP
A sample or group of samples that are included in an evolutionary tree because they are known, or assumed, to connect directly to the root of the tree (that is, to the node of the tree that represents the common ancestor of all samples in the tree).

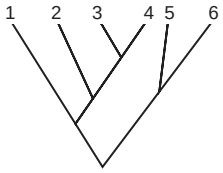
HOMOPLASY
Identical character states (for example, the same nucleotide base in a DNA sequence) that are not the result of common ancestry (not homologous), but arose independently in different ancestors by parallel or convergent mutations.

LINKAGE BLOCK
A region of DNA that is inherited as a single unit owing to a lack of recombination, such as the mitochondrial DNA of metazoans. The histories of genes that are located in such regions are not independent, and are equally affected by all the selective forces that have acted anywhere in the linkage block.

a

Outgroup	TGTACGTACCG
Sequence 1	.C.....G.T.
Sequence 2	...G.C.G...
Sequence 3	..C..C.G...
Sequence 4	..C..C.G..A
Sequence 5	C...T.C.C..
Sequence 6	....T...C..

b



c

$$\hat{\theta} = S / \sum_{i=1}^{n-1} \frac{1}{i} = \frac{11}{2.283} = 4.8$$

Figure 2 | **Contrasting phylogeographic and summary-statistic methods.** Phylogeographic and summary-statistic approaches to population-genetic analysis are represented by their typical starting points, beginning with a data set of polymorphic sites in a DNA sequence. **a** | A data set of six sequences, showing only the variable positions together with the outgroup sequence. **b** | Phylogeographic studies begin with a gene-tree estimate that is based on the data. **c** | Many population-genetic studies that take a summary-statistic approach begin by estimating the fundamental population-genetic parameter θ, assuming a Wright–Fisher population. Watterson's (1975) estimate of the population mutation rate ($\theta = 4Nu$) is given by the formula shown, in which the number of polymorphic sites (S) is 11 and the sample size (n) is 6 and u is the neutral mutation rate[44].

attracted much interest, as it seems to offer the promise of detailed findings of the sort that have usually eluded research on population structure. The INFERENCE-KEY component of NCA is essentially a summary of the kinds of patterns that would be expected in a network under different models of the causes of geographic structure in the data[30]. Although, in some respects, they are quantitative, the expectations that are represented in the inference key are not the result of analytical models, and the inferences that result are not statistical in nature. The method does not provide assessments of confidence for any particular interpretation, and it is not known how frequently an interpretation is incorrect — an NCA assessment of a particular factor in the history of a data set could, in fact, be the result of a history that does not include that factor. The NCA method has been criticized on these grounds, and simulations have shown that the inference key can lead to conclusions of historical processes that did not actually occur[33].

Limitations of gene-tree-based methods

Gene trees are appealing because they are graphically accessible and not explicitly tied to any particular historical model or previous idea of the locations and boundaries of populations. Investigators are free to take gene-tree patterns at face value, without preconceptions about the models, and to follow a HEURISTIC path of interpretation. By contrast, investigators that rely on explicit population models, such as those in FIG. 1, often face the difficulty at the outset of not knowing what kind of model to apply.

The most pervasive and general difficulty with gene-tree interpretation is that most demographic histories (including the simple Wright–Fisher model) lead to gene trees that vary widely in their topology and branch lengths. This variance, often referred to as STOCHASTIC VARIANCE in a population-genetic context, is the variation in actual genealogical histories that are experienced by different genes, which cannot be reduced by increasing the sample size for a gene (either by increasing the number or length of DNA sequences). FIG. 3 shows how a wide variation in branching pattern can occur, with different genes seeming to have different histories even if they come from the same populations. The large stochastic variance among genes means that any individual gene tree (for example, based on mtDNA) will probably be far removed from any estimate of history that would be obtained using a large number of genes. This means that a study based on one gene is usually not sufficient to determine whether or not two populations are exchanging genes. Similarly, the depth of a gene tree that includes samples from two populations or species is often a poor indicator of when those populations began to diverge[34]. Also, the overall depth of a gene tree cannot be used to accurately predict what might be found at another gene[35].

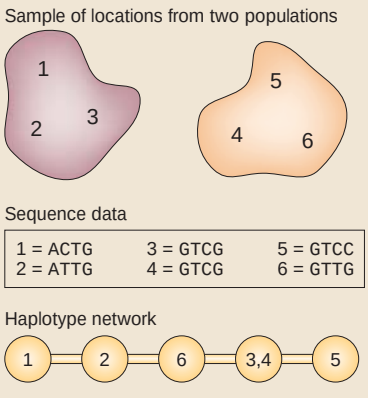
ALLOPATRIC DIFFERENTIATION
The process of divergence between populations or species that are geographically separated.

INFERENCE KEY
A list of paired rules that are used for diagnosis or identification. Keys are a classic tool for identifying organisms to the species level, on the basis of the presence or absence of specific morphological characters or character states. A similar tool is used in nested-clade analysis to distinguish between different historical scenarios.

HEURISTIC
A method of inference that relies on educated guesses or simplifications that limit the parameter space over which solutions are searched. This approach is not guaranteed to find the correct answer.

STOCHASTIC VARIANCE
In the context of gene histories, this is the variation in gene trees and mutations among unlinked genes that have passed through the same demographic history of populations of organisms.

Box 2 | Nested-clade analysis



For a set of DNA sequences that are sampled from one or more populations, the first step in nested-clade analysis (NCA) is to estimate the haplotype network, which is similar to a gene-tree estimate. The difference is that in a network, sampled haplotypes can be found at the internodes as well as the tips[94]. Each step along the network represents one mutation change that was observed in the data. Next, nested groups of haplotypes (clades) that are separated from one another by one or more mutational steps are identified[95,96]. A test of geographic structure determines whether samples from the same population are closer to each other in the haplotype network than would be expected by chance. Finally, if evidence of geographic structure is found, the relationships between mutational distance and geographical distance among haplotypes are interpreted using an inference key of possible demographic factors[29,30]. The test and subsequent inferences rely on two calculated geographical distances: D_c (clade distance) and D_n (nested-clade distance). D_c measures the geographical range of a given clade, and corresponds to the average distance of each member of a clade from its geographic centre. D_n measures the average geographic distance of all members of a clade from the geographical centre of its higher-level nesting clade, which is also estimated by averaging the coordinates of all members of the higher-level nesting clade. The inference key was constructed on the basis of expected patterns of geographical association that can arise under three types of historical event: restricted gene flow, range expansion and allopatric fragmentation. The program GeoDis can be used to conduct NCA[97].

MONOPHYLY
The property that is attributed to a group of samples in an evolutionary tree that all share the same common ancestor exclusive of other samples in the tree. A set of samples that constitute an entire branch on an evolutionary tree is said to be monophyletic.

Stochastic variance is also the reason why MONOPHYLY of populations in gene trees is often a poor guide for species status[36]. As divergence begins, monophyly arises slowly and with wide variation among genes[36].

Another difficulty with gene-tree approaches that rely on a single locus is that the history of any one gene might have been strongly affected by natural selection and, therefore, might not reflect the demographic history of the sampled populations. In particular, genes that have experienced the selective replacement of beneficial mutations will have gene-tree histories that are short, with the common ancestor being a gene copy that was among the first to carry the beneficial mutation. mtDNA and other completely linked blocks of genes are of concern in this regard because a beneficial mutation at any one of the genes will alter the gene-tree history for the others[37].

Tree-based methods are also inherently limited by the accuracy of gene-tree estimates. Many mitochondrial data sets come from histories with large amounts of recurrent mutation, so it is difficult to be sure that the tree estimate (or corresponding network) accurately

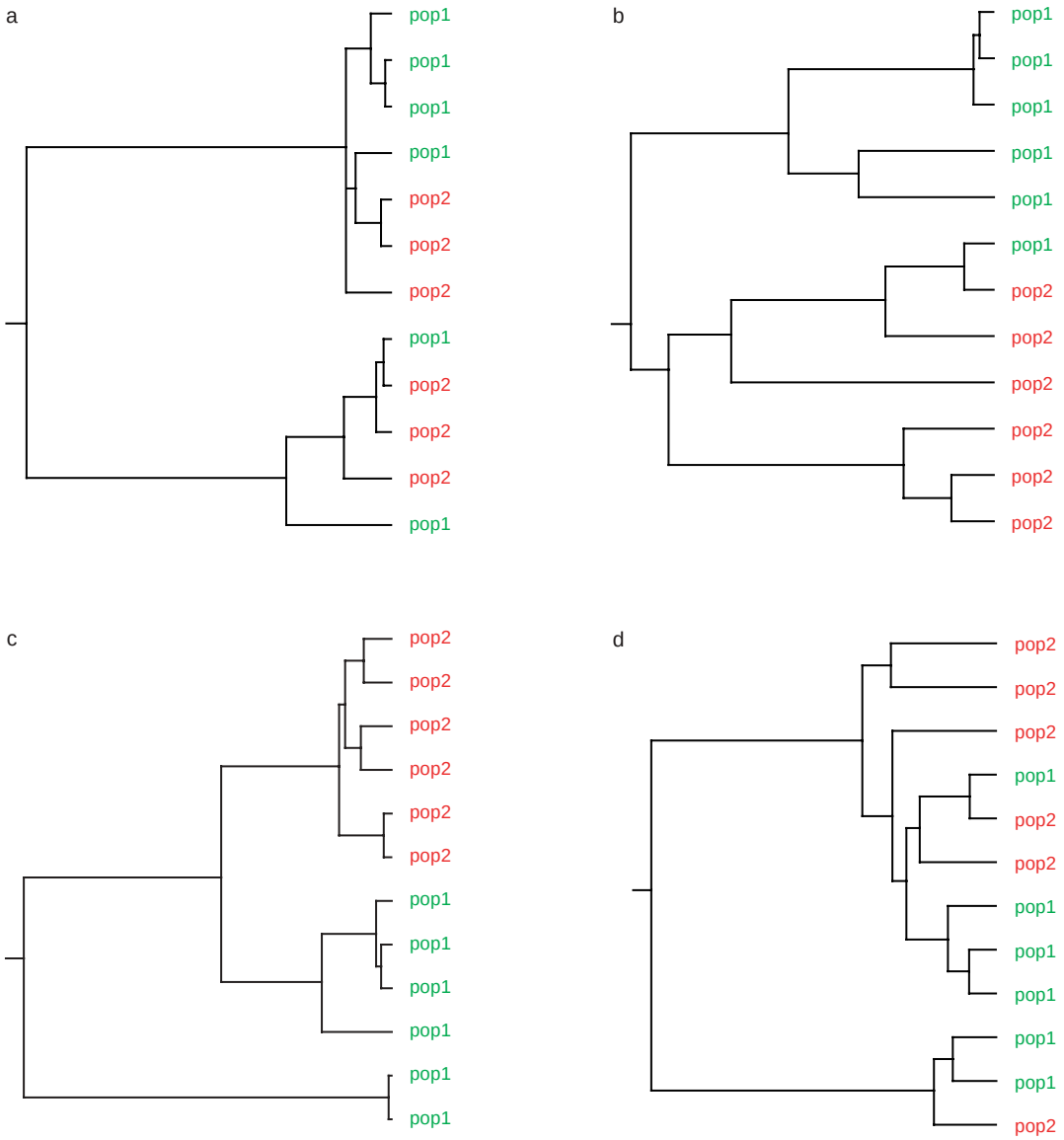


Figure 3 | **The stochastic variance of gene trees.** Coalescent simulations were done with six gene copies that were sampled from two populations —'pop1' (green) and 'pop2' (red) — each with 2*N* gene copies. The populations split from an ancestral population 4*N* generations ago and gene flow was simulated after separation at a rate of 0.5 migrants per generation in each direction. Six simulations were done, and four are shown to illustrate the various histories that are indicated by branching patterns and branch lengths. **a** | The tree indicates the presence of two long-separated populations that might have recently exchanged gene copies so that some pop1 sequences cluster with pop2 sequences and *vice versa*. **b** | The tree indicates a history of two diverging populations with a single instance of recent gene exchange from pop2 to pop1 (yielding a gene copy that is identified as pop1 on the basis of sampling, but that clusters with pop2 sequences). **c** | The tree seems to indicate a history in which pop2 separated from pop1 some time ago, without subsequent gene exchange. **d** | The tree resembles the pattern that might be expected if pop1 and pop2 were in fact a single population, as pop1 and pop2 sequences are intermingled.

represents the true history. To the extent that the tree estimate is wrong, the conclusions that depend on the tree will also be wrong. In general, this problem is well appreciated because of the attention it has received in systematics contexts (in which trees are often deep and have histories with many mutations), and there is a large body of literature on the difficulties of tree estimation and on how to assess confidence in tree estimates[38,39]. A related problem is that many data sets from nuclear genes have histories that include recombination, which by their nature cannot be represented by a branching diagram[40]. If recombination has not been too frequent, some recombination events can be reconstructed from the pattern of variation, and this information can be incorporated into a phylogeographic analysis[41].

Summary statistics — methods and applications

Methods that take a summary-statistic approach to DNA sequence variation have largely grown out of neutral models (that is, models that do not include natural selection) that were developed before the advent of DNA sequence-based studies[12,42–44]. Much of the new theory is directly connected to models that existed before the coalescent age, and there are many contexts in which allelic models and DNA sequence models are directly interchangeable. For example, Wright's classic F-STATISTIC indices of population structure[45,46] are directly translatable to a DNA sequence context[47,48]. Similarly, the different equilibrium models that are shown in FIG. 1 have been examined using coalescent methods[49–55].

The most commonly used summary statistics focus on the theoretical idea of a population mutation rate, typically denoted by θ, which is equal to $4Nu$, where N is the population size and u is the neutral mutation rate for the gene in question (FIG. 2). In a Wright–Fisher population, the expected time to the common ancestor of two randomly selected sequences is $2N$ generations. So, the number of neutral mutations that separate the two sequences is expected to be $2N \times u$, multiplied by a further factor of two because each descendant gene copy has experienced $2N$ generations, on average, since the common ancestor. The parameter can be estimated using either the number of polymorphic sites in the sample (S)[44], the average number of pairwise differences between the sampled sequences (k)[56], or the number of SINGLETON MUTATIONS (η)[57].

When basic summary-statistics methods were first applied to studies of human populations, it quickly became clear that traditional equilibrium models were not appropriate. For one thing, it was known that human populations have grown, which violates a basic assumption in the Wright–Fisher model. In this regard one of the first non-equilibrium models to receive considerable attention addressed population-size change. It had been shown that although both S and k can be affected by changes in population size, the effects on k are stronger. So, one way to assess whether data fit a constant-size population model is to see if two estimates of θ, one based on S and the other on k,

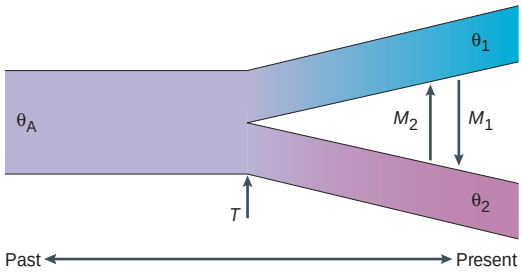


Figure 4 | **The isolation-with-migration model.** This model assumes that an ancestral population of constant size and population mutation parameter θ_A separated into two populations at time T. Each descendant population has its own mutation parameter, θ_1 and θ_2, respectively. Also, migration occurs between the two populations at rates M_1 and M_2. This is a general model that includes, as a special case, the isolation-without-migration model (by setting the migration rates to zero).

are compatible with each other. A statistic based on the difference between the two — Tajima's D[58] — has been widely used to test for changes in population size. Other approaches have been developed to assess departures from population-size constancy[5,59–61]. In recent years, coalescent-based summary-statistic approaches have also been applied to non-equilibrium models[5,6]. Alternative summary-statistics methods have been used to estimate the time of divergence among populations, the rate of migration among populations[62] and the effect of population structure on DNA polymorphism[34,50,63–68].

Limitations of summary-statistics methods

The use of summary statistics necessarily goes hand-in-hand with the adoption of a particular demographic model, such as an island model or an isolation by distance model (FIG. 1). It is the model that specifies the meaning of the parameters and the assumptions that underlie them. This does not mean that investigators must assume the model to be correct, as there are ways to use the summary statistics to assess the fit between data and a model[69]. It does mean, however, that the scope of demographic models that can be explored by an investigator is limited by the availability of statistics for those models. By contrast, phylogeographic approaches are not explicitly model bound.

Another problem with summary statistics is that they do not necessarily take advantage of all of the information in the data. For example, the simple summary statistic S (the number of variable positions in the DNA sequence) is directly connected to θ (assuming a Wright–Fisher model), but samples from two different populations might have a common S despite having widely different demographic histories. By itself, S says little about history. One approach to overcome this limitation is to summarize the pattern of variation using several statistics. For example, the POLYMORPHIC-SITE FREQUENCY DISTRIBUTION is a series of counts that captures much of the information in a DNA sequence data set[70,71]. It is also possible to select a

F-STATISTICS
A method of summary statistics that was devised by Sewall Wright to describe correlations among alleles that are sampled at different hierarchical levels (individuals, subpopulations and total populations). F-statistics are frequently used to describe the presence of population structure.

SINGLETON MUTATIONS
Polymorphic sites in which a rare base is found in only one of the sampled sequences.

set of summary statistics so as to optimize the capture of information and allow estimates of model parameters that are of nearly the quality that would be found using all of the information in the data[72,73].

Summary statistics offer no panacea for the wide stochastic variance that occurs among genes, and conclusions that are based on a single gene are likely to be misleading because of this variance or the effects of natural selection, as is the case in a phylogeographic context. However, an important difference is that with a mathematical approach it is sometimes possible to estimate the stochastic variance that is associated with the summary statistics, or at least to estimate the overall variance, which includes both the stochastic and the sampling components of the variance[44,56,74].

Bridging the divide

The study of evolution in structured populations has been divided between a gene-tree-based phylogeographic approach and the more traditional mathematical approach that relies on explicit models and summary statistics. What is needed are methods that have the advantages of both approaches, that are not tightly bound to specific assumption-laden models and that take full advantage of the data, as is the case with tree-based methods, and yet provide the quantitative and statistical rigour of the summary-statistics methods[33].

There are some methods for which estimates of model parameters are based on gene-tree estimates. These hybrid approaches begin with a tree estimate and then proceed to estimate model parameters from features of the tree. As with gene-tree approaches in general, these methods must usually overlook the uncertainty that is associated with the gene tree, including the possibility that the tree topology is incorrect, as well as the stochastic variance among trees that are sampled from the same population. For example, Slatkin and Maddison[75] developed a method for estimating the population migration rate between two populations, using the branching pattern of an estimated gene tree. A different group of methods use gene trees but focus on the branch lengths between the nodes of the tree, rather than on the branching pattern, to detect changes in the effective population size (*Ne*)[76–79].

In recent years, population geneticists have begun to develop new approaches that in some ways unite the tree-based and summary-statistic-based methods[14,80]. There are two main limitations of phylogeographic methods: the possibility that a gene-tree estimate differs from the true tree and the wide stochastic variance among the trees of independent genes. To see this in a rough mathematical framework and to introduce the newer family of methods, consider an estimated gene tree G that is based on a data set D. Suppose that we would like to calculate the value of a function F of that tree, which is based on D and assumes a particular historical model Θ — that is, $F(G \mid D, \Theta)$. One kind of function that is often useful provides an estimate of a parameter in the model. The difficulty is that the value of the function will vary with the gene tree G, and therefore the value of $F(G \mid D, \Theta)$ for any particular gene tree will usually not be useful because of the large stochastic variance that is associated with G. This is true even if we use a tree that is the best possible estimate of the true tree, given the data. What we would like to do is calculate a related function that takes advantage of all of the information in the data, but does not depend on a particular tree — that is, $F(D \mid \Theta)$. What is needed is a way to consider all of the many possible gene trees that are consistent with the data. The newer family of methods can do this, as they make use of coalescent calculations of the probability of obtaining a particular gene-tree estimate as a function of a particular historical model. Let this probability be denoted by $P(G \mid \Theta)$. We can then consider all of the possible gene trees, and weight each by their probability, to calculate the function $F(D \mid \Theta) = \Sigma_G F(G \mid D, \Theta) P(G \mid \Theta)$.

In practice, a computer program (such as MDIV, Batwing, Genetree or LAMARC) is used to carry out these calculations and to handle the large number of gene trees. Fortunately, the summation usually need not include all possible trees as long as the sample of trees is a random collection of those that are possible for a given data set. So, we might not know the value of $F(D \mid \Theta)$ precisely, but we might still be able to obtain a good estimate of it. In the most complete implementations, $F(D \mid \Theta)$ is the likelihood of a set of model parameter values for a given tree, and we can estimate complete likelihood or probability distributions for the model parameters. These methods all take full account of the inherent stochastic variance of gene trees by considering all (or a great many) possible trees and weighting them in proportion to how likely they are given the data and the model. Such methods have been used to estimate population mutation rate[81–83], population growth rate[84,85], population recombination rates[86,87] and migration rates[84,88]. These probability-based methods take advantage of all of the information in the data and, increasingly, can be applied to complex models. One of the most complete implementations concerns a demographic model that is of wide interest for the study of diverging populations[89–91] (FIG. 4). The 'isolation-with-migration' model has many parameters (six) compared with most other methods, which have only one or two, yet the method for estimating the parameters is able to generate complete probability distributions for all of them[92].

The drawback of these methods is that they are complex and difficult to implement. Also, they have not yet been developed to the extent that they can be applied in an accessible exploratory manner that mirrors the heuristic thought process that often accompanies the examination of individual gene-tree estimates. However, these methods are being developed for increasingly complex models, and, in principle, it might be possible to develop general computer-based tools that offer a wide array of potential models, together with statistical methods to choose among them. The methods can also be applied simultaneously to data from multiple loci,

POLYMORPHIC-SITE FREQUENCY DISTRIBUTION
A polymorphic site in a DNA sequence can be described by the frequency of one of its variable bases. The distribution of these values for all the polymorphic sites in a sample can be described using a histogram or bar chart. The shape of the histogram can provide qualitative information on the processes that are involved in the history of the sample.

particularly if those loci are effectively unlinked so that their histories are independent in the model. As larger data sets become available for increasingly large portions of the genome, these methods should be able to grow with them. The hope is that in the future, investigators will be able to take full advantage of all of the information in the data to explore a wide variety of historical models in a way that allows assessments of how likely different models are, and allows estimates of the components (parameters) in different models.

1. Provine, W. B. *The Origins of Theoretical Population Genetics* (Univ. of Chicago Press, Chicago, 1971).
2. Fisher, R. *The Genetical Theory of Natural Selection* (Clarenson, Oxford, 1930).
3. Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97–159 (1931).
   The first paper to mathematically address the effects of population structure on patterns of genetic variation.
4. Wright, S. *Evolution and the Genetics of Populations Volume 2: The Theory of Gene Frequencies* (Univ. of Chicago Press, Chicago, 1969).
5. Wakeley, J. & Hey, J. Estimating ancestral population parameters. *Genetics* **145**, 847–855 (1997).
6. Wakeley, J. Nonequilibrium migration in human history. *Genetics* **153**, 1863–1871 (1999).
7. Slatkin, M. Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* **47**, 264–279 (1993).
8. Van Dooren, T. J. M. & Metz, J. A. J. Delayed maturation in temporally structured populations with non-equilibrium dynamics. *J. Evol. Biol.* **11**, 41–62 (1998).
9. Avise, J. C. *et al.* Intraspecific phylogeography: the mitochondrial-DNA bridge between population genetics and systematics. *Annu. Rev. Ecol. Syst.* **18**, 489–522 (1987).
   This review paper marks the birth of phylogeography.
10. Avise, J. C. *Phylogeography* (Harvard Univ. Press, Cambridge, Massachusetts, 2000).
11. Bermingham, E. & Mortiz, C. Comparative phylogeography: concepts and applications. *Mol. Evol.* **7**, 367–369 (1998).
12. Kingman, J. F. C. The coalescent. *Stoch. Proc. Appl.* **13**, 235–248 (1982).
    The original mathematical description of the coalescent theory.
13. Hudson, R. R. in *Oxford Surveys in Evolutionary Biology* (eds Futuyma, D. & Antonovics, J.) 1–44 (Oxford Univ. Press, New York, 1990).
    A comprehensive review of coalescent theory by one of its developers, which provides computer code for conducting basic simulations of neutral processes.
14. Rosenberg, N. A. & Nordborg, M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Rev. Genet.* **3**, 380–390 (2002).
15. Tavare, S. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.* **26**, 119–164 (1984).
16. Hare, M. P. Prospects for nuclear gene phylogeography. *Trends Ecol. Evol.* **16**, 700–706 (2001).
17. Bernardi, G., Sordino, P. & Powers, D. A. Concordant mitochondrial and nuclear DNA phylogenies for populations of the teleost fish *Fundulus heteroclitus*. *Proc. Natl Acad. Sci. USA* **90**, 9271–9274 (1993).
18. Burton, R. S. & Lee, B. N. Nuclear and mitochondrial gene genealogies and allozyme polymorphism across a major phylogeographic break in the copepod *Tigriopus californicus*. *Proc. Natl Acad. Sci. USA* **91**, 5197–5201 (1994).
19. Palumbi, S. R. & Baker, C. S. Contrasting population structure from nuclear intron sequences and mtDNA of humpback whales. *Mol. Biol. Evol.* **11**, 426–435 (1994).
20. Hare, M. P. & Avise, J. C. Population structure in the American oyster as inferred by nuclear gene genealogies. *Mol. Phylogenet. Evol.* **15**, 119–128 (1998).
21. Hare, M. P., Cipriano, F. & Palumbi, S. R. Genetic evidence on the demography of speciation in allopatric dolphin species. *Evolution* **56**, 804–816 (2002).
22. Machado, C. A. & Hey, J. The causes of phylogenetic conflict in a classic *Drosophila* species group. *Proc. Royal Soc. Lond. B* **270,** 1193–1202 (2003).
23. Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).
    A much-discussed paper that describes one of the first attempts to use mitochondrial DNA data to study the history of the human species.
24. Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503–1507 (1991).
25. Maddison, D. R., Ruvolo, M. & Swofford, D. L. Geographic origins of human mitochondrial DNA: phylogenetic evidence from control region sequences. *Syst. Biol.* **41**, 111–124 (1992).
26. Templeton, A. R. Human origins and analysis of mitochondrial DNA sequences. *Science* **255**, 737 (1992).
27. Templeton, A. R. The "Eve" hypothesis: a genetic critique and reanalysis. *Am. Anthropol.* **95**, 51–72 (1993).
28. Hey, J. Mitochondrial and nuclear genes present conflicting portraits of human origins. *Mol. Biol. Evol.* **14**, 166–172 (1997).
29. Templeton, A. R., Routman, E. & Phillips, C. A. Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics* **140**, 767–782 (1995).
    The original description of the nested-clade-analysis method.
30. Templeton, A. R. Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Mol. Ecol.* **7**, 381–397 (1998).
31. Templeton, A. Out of Africa again and again. *Nature* **416**, 45–51 (2002).
32. Stringer, C. B. & Andrews, P. Genetic and fossil evidence for the origins of modern humans. *Science* **239**, 1263–1268 (1988).
33. Knowles, L. L. & Maddison, W. P. Statistical phylogeography. *Mol. Ecol.* **11**, 2623–2635 (2002).
34. Edwards, S. V. & Beerli, P. Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* **54**, 1839–1854 (2000).
35. Hudson, R. R. & Turelli, M. Stochasticity overrules the "three-times rule": genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. *Evolution* **57**, 182–190 (2003).
36. Hudson, R. R. & Coyne, J. A. Mathematical consequences of the genealogical species concept. *Evolution* **56**, 1557–1565 (2002).
37. Maynard Smith, J. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genome Res.* **23**, 23–35 (1974).
38. Felsenstein, J. Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Gen.* **22**, 521–565 (1988).
39. Swofford, D., Olsen, G., Waddel, P. & Hillis, D. in *Molecular Systematics* (eds. Hillis, D., Mortiz, C. & Mable, B.) 486–493 (Sinauer Associates, Sunderland, Massachusetts, 1996).
40. Hudson, R. R. & Kaplan, N. L. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164 (1985).
41. Templeton, A. R. *et al.* Recombinational and mutational hotspots within the human lipoprotein lipase gene. *Am. J. Hum. Genet.* **66**, 69–83 (2000).
42. Kimura, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893–903 (1969).
43. Ewens, W. J. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3**, 87–112 (1972).
44. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–275 (1975).
45. Wright, S. The genetical structure of populations. *Ann. Eugen.* **15**, 323–354 (1951).
46. Wright, S. The interpretation of population structure by F-statistics with special regards to systems of mating. *Evolution* **19**, 395–420 (1965).
47. Slatkin, M. & Voelm, L. Fst in a hierarchical island model. *Genetics* **127**, 627–629 (1991).
48. Slatkin, M. Inbreeding coefficients and coalescence times. *Genome Res.* **58**, 167 (1991).
49. Notohara, M. The coalescent and the genealogical process in geographically structured population. *J. Math. Biol.* **29**, 59–75 (1990).
50. Wakeley, J. Segregating sites in Wright's Island model. *Theor. Popul. Biol.* **53**, 166–174 (1998).
51. Wakeley, J. The effects of subdivision on the genetic divergence of populations and species. *Evolution* **54**, 1092–1101 (2000).
52. Wilkins, J. F. & Wakeley, J. The coalescent in a continuous, finite, linear population. *Genetics* **161**, 873–888 (2002).
53. Whitlock, M. C. Neutral additive genetic variance in a metapopulation. *Genet. Res.* **74**, 215–221 (1999).
54. Wakeley, J. & Aliacar, N. Gene genealogies in a metapopulation. *Genetics* **159**, 893–905 (2001).
55. Hey, J. A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. *Theor. Popul. Biol.* **39**, 30–48 (1991).
56. Tajima, F. Evolutionary relationships of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
57. Fu, Y. X. Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. *Genetics* **138**, 1375–1386 (1994).
58. Tajima, F. The effect of change in population size on DNA polymorphism. *Genetics* **123**, 597–601 (1989).
59. Slatkin, M. & Hudson, R. R. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**, 555–562 (1991).
60. Rogers, A. R. & Harpending, H. Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* **9**, 552–568 (1992).
61. Innan, H. & Stephan, W. The coalescent in an exponentially growing metapopulation and its application to *Arabidopsis thaliana*. *Genetics* **155**, 2015–2019 (2000).
62. Hudson, R. R., Slatkin, M. & Maddison, W. P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589 (1992).
63. Tajima, F. DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-subpopulation model. *Genetics* **123**, 229–240 (1989).
64. Tajima, F. Relationship between migration and DNA polymorphism in a local population. *Genetics* **126**, 231–234 (1990).
65. Slatkin, M. The average number of sites separating DNA sequences drawn from a subdivided population. *Theor. Popul. Biol.* **32**, 42–49 (1987).
66. Strobeck, C. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics* **117**, 149–153 (1987).
67. Wakeley, J. Pairwise differences under a general model of population subdivision. *J. Genet.* **75**, 81–89 (1996).
68. Arbogast, B. S., Edwards, S. V., Wakeley, J., Beerli, P. & Slowinski, J. B. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annu. Rev. Ecol. Syst.* **33**, 707–740 (2002).
69. Ford, M. J. Applications of selective neutrality tests to molecular ecology. *Mol. Ecol.* **11**, 1245–1262 (2002).
70. Braverman, J. M., Hudson, R. R. & Stephan, W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140**, 783–796 (1990).
71. Fu, Y. X. & Li, W. H. Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709 (1993).
72. Tavare, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518 (1997).
73. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002).
74. Hudson, R. R., Kreitman, M. & Aguadé, M. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159 (1987).
75. Slatkin, M. & Maddison, W. P. A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics* **123**, 603–613 (1989).
    The first method that was developed to estimate migration rates using a gene tree.
76. Felsenstein, J. Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method. *Gene. Res.* **60**, 209–220 (1992).
    The first study to describe a method to estimate a population-genetic parameter (population size) by integrating over multiple gene trees.
77. Fu, Y. X. A phylogenetic estimator of effective population size or mutation rate. *Genetics* **136**, 685–692 (1994).
78. Nee, S., Holmes, E. C., Rambaut, A. & Harvey, P. H. Inferring population history from molecular phylogenies. *Phil. Trans. Royal Soc. Lond. B* **349**, 25–31 (1995).
79. Pybus, O. G., Rambaut, A. & Harvey, P. H. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* **155**, 1429–1437 (2000).
80. Felsenstein, J., Kuhner, M. K., Yamato, J. & Beerli, P. in *Statistics in Genetics and Molecular Biology* (ed. Seillier-Moiseiwitsch, F.) (Institute of Mathematical Statistics and American Mathematical Soc., Hayward, California, 1999).

81. Griffiths, R. C. & Tavare, S. Simulating probability distributions in the coalescent. *Theor. Popul. Biol.* **46**, 131–159 (1994).
82. Griffiths, R. C. & Tavare, S. The age of a mutation in a general coalescent tree. *Stochastic Models* **14**, 273–295 (1998).
83. Kuhner, M. K., Yamato, J. & Felsenstein, J. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140**, 1421–1430 (1995).
84. Bahlo, M. & Griffiths, R. C. Inference from gene trees in a subdivided population. *Theor. Popul. Biol.* **57**, 79–95 (2000).
85. Kuhner, M. K., Yamato, J. & Felsenstein, J. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics.* **149**, 429–434 (1998).
86. Kuhner, M. K., Yamato, J. & Felsenstein, J. Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**, 1393–1401 (2000).
87. Nielsen, R. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**, 931–942 (2000).
88. Beerli, P. & Felsenstein, J. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152**, 763–773 (1999).
89. Takahata, N. & Slatkin, M. Genealogy of neutral genes in two partially isolated populations. *Theor. Popul. Biol.* **38**, 331–350 (1990).
The first paper to address the difficulty of distinguishing the presence of gene flow in a non-equilibrium isolation model.
90. Hey, J. in *Molecular Approaches to Ecology and Evolution* (eds. Schierwater, B., Streit, B., Wagner, G. & DeSalle, R.) 435–449 (Birkhäuser, Basel, 1994).
91. Wakeley, J. & Hey, J. in *Molecular Approaches to Ecology and Evolution* (eds. DeSalle, R. & Schierwater, B.) 157–175 (Birkhäuser, Basel, 1998).
92. Nielsen, R. & Wakeley, J. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**, 885–896 (2001).
93. Moran, P. A. P. Random processes in genetics. *Camb. Philos. Soc. Proc.* **54**, 60–71 (1958).
94. Templeton, A. R., Crandall, K. A. & Sing, C. F. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimating. *Genetics* **132**, 619–633 (1992).
95. Templeton, A. R., Boerwinkle, E. & Sing, C. F. Cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. *Genetics* **117**, 343–351 (1987).
96. Templeton, A. R. & Sing, C. F. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping IV. Nested analyses with cladogram uncertainty and recombination. *Genetics* **134**, 659–669 (1993).
97. Posada, D., Crandall, K. A. & Templeton, A. R. GeoDis: a program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. *Mol. Ecol.* **9**, 487–488 (2000).
98. Wright, S. Breeding structure of populations in relation to speciation. *Am. Nat.* **74**, 232–248 (1940).
99. Kimura, M. & Weiss, G. H. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49**, 561–576 (1964).
100. Wright, S. Isolation by distance. *Genetics* **28**, 114–138 (1943).
101. Malecot, G. *The Mathematics of Heredity* (Freeman, San Francisco, 1969).
102. Slatkin, M. Gene flow and genetic drift in a species subject to frequent local extinction. *Theor. Popul. Biol.* **12**, 253–262 (1977).
103. Wade, M. J. & McCauley, D. E. Extinction and recolonization: their effects on the genetic differentiation of local populations. *Evolution* **42**, 995–1005 (1988).

Acknowledgements

We are grateful to M. Hare, Y.-J. Won and two anonymous referees for helpful suggestions and corrections. This work was supported in part by a grant from the National Institutes of Health to J.H.

Online links

FURTHER INFORMATION
Batwing: http://www.maths.abdn.ac.uk/~ijw/downloads/download.htm
Genetree: http://www.stats.ox.ac.uk/~griff/software.html
GeoDis: http://inbio.byu.edu/Faculty/kac/crandall_lab/geodis.htm
LAMARC: http://evolution.genetics.washington.edu/lamarc.html
MDIV: http://www.biom.cornell.edu/Homepages/Rasmus_Nielsen/files.html
Access to this interactive links box is free online.