GENEALOGICAL TREES, COALESCENT THEORY AND THE ANALYSIS OF GENETIC POLYMORPHISMS

Noah A. Rosenberg and Magnus Nordborg

Improvements in genotyping technologies have led to the increased use of genetic polymorphism for inference about population phenomena, such as migration and selection. Such inference presents a challenge, because polymorphism data reflect a unique, complex, non-repeatable evolutionary history. Traditional analysis methods do not take this into account. A stochastic process known as the 'coalescent' presents a coherent statistical framework for analysis of genetic polymorphisms.

POLYMORPHISM DATA
Data that include the genotypes of many individuals sampled at one or more loci; here we consider a locus to be polymorphic if two or more distinct types are observed, regardless of their frequencies.

HAPLOTYPE
The allelic configuration of multiple genetic markers that are present on a single chromosome of a given individual.

COALESCENCE
The merging of ancestral lineages going back in time.

STOCHASTIC PROCESS
A mathematical description of the random evolution of a quantity through time.

Program in Molecular and Computational Biology, University of Southern California, 835 West 37th Street, SHS172, Los Angeles, California 90089-1340, USA. Correspondence to M.N. e-mail: magnus@usc.edu
DOI: 10.1038/nrg795

In their classic experiment, Luria and Delbrück[1] observed independent runs of the evolution of phage resistance in bacterial populations that were initially susceptible to phage infection. For each run, the frequency of phage resistance at the end of the experiment was measured. The goal was to test hypotheses about the processes that gave rise to variation: the frequency of phage resistance varied considerably across the runs, in a manner more consistent with the spontaneous, random occurrence of mutations than with mutation to phage resistance after exposure to phage.

It is illuminating to compare the Luria–Delbrück experiment with the efforts of modern evolutionary geneticists to "make sense out of sequence"[2]. Like Luria and Delbrück, we seek to use POLYMORPHISM DATA to test models about evolutionary forces (although we might be more interested in historical demography or selection than in mutation). However, unlike them, we analyse polymorphisms collected from natural populations. This modern approach leads to two fundamental problems — first, because there is no replication of the 'experiment', only one run of evolution is available to be studied, and second, the starting conditions of the 'experiment' are unknown. These problems might seem obvious, but it is not always appreciated how profound their implications are for data analysis. Consider a sample of HAPLOTYPES from a population. For each haplotype, the allelic states of the different loci are statistically dependent owing to genetic linkage, and for each locus, the allelic states of different haplotypes are statistically dependent owing to their shared ancestry. These dependencies are the result of the unique history of mutation, recombination and COALESCENCE of lineages in the ancestry of the sample. These facts must be incorporated if the data are to be analysed in a coherent statistical framework. Heuristic methods, such as those borrowed from phylogenetics, do not fully take into account the uncertainty caused by the inherent randomness of evolution, and as a result can lead to pronounced overinterpretations of the data.

One solution is to model the past using a suitable stochastic model. The STOCHASTIC PROCESS known as 'the coalescent' is a natural extension of classical population-genetics models and is very well suited for this purpose. It is relatively simple and can be adapted to accommodate a wide variety of biological assumptions. In this review, we describe the coalescent and how it can be used to analyse data. Surveys of recent developments in this field are available elsewhere[3–8]. We begin by focusing on why history must be considered when analysing polymorphism data.

Importance of history

Any model of DNA polymorphism in a population must include mutation — without which there would

BACTERIAL CONJUGATION
Genetic recombination in prokaryotes that is mediated through direct transfer of DNA from a donor to a recipient cell.

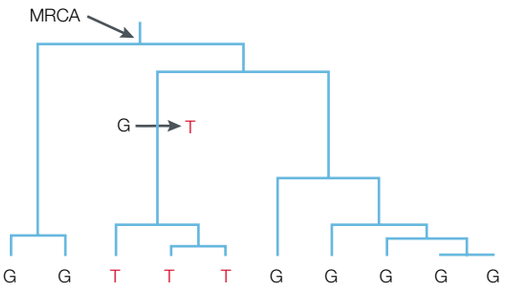


Figure 1 | **The source of genetic variation.** Polymorphism at a particular site results from mutations (shown here as G→T) along branches of the genealogical tree, which connects sampled copies of the site to their most recent common ancestor (MRCA).

be no polymorphism — as well as the genealogy of sampled sequences. To model the genealogy, we need to consider the recombination and coalescence of lineages.

Coalescence and mutation. Consider a particular site in the genome of a species. All existing copies of this site must be related to each other and to a most recent common ancestor (MRCA) through some form of genealogical tree. Polymorphism at the site is due to mutations that occurred along the branches of this tree, and the frequency of each sequence variant is determined by the fraction of branches that inherits the variant (FIG. 1). The pattern of polymorphism therefore reflects both the history of the coalescence of lineages, which gives rise to the tree, and the mutational history.

To observe the effect of history on data analysis, imagine that we sequence a 10-kb region in 30 randomly chosen individuals and, surprisingly, find no polymorphisms. We might interpret this observation as evidence for selective constraint in this region. Alternatively, it might be that the individuals chosen for the comparison are unusually closely related. So, the interpretation depends on the genealogy of the sequences, which is not known.

To deal with this uncertainty, we treat the genealogy as random, in the same way that we treat mutation as random. Just as mutations occur differently across runs of evolution[1], if evolution were repeated, samples from different 'runs' of evolution would have different genealogical trees (FIG. 2). It is necessary to incorporate both of these sources of variation into data analysis — the randomness of genealogies and of mutations. For example, to decide if the data in the previous paragraph are unusual, we might make assumptions about the process that gave rise to those data, and imagine many random repetitions of the evolutionary process. If the fraction of the random genealogical and mutational histories that could have given rise to the observed data is small, we can conclude that the assumptions cannot explain the pattern. To consider genealogies that might be found in different runs of evolution, we need models that allow us to construct random genealogies, and the coalescent is one such model.

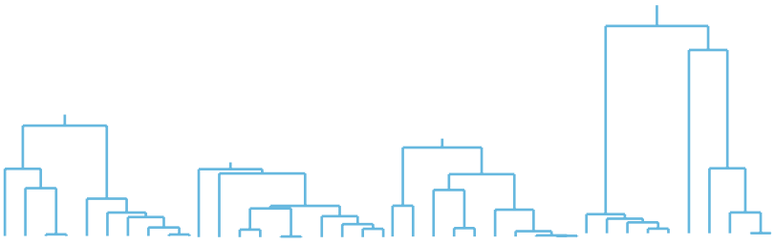


Figure 2 | **Random genealogical trees.** The trees were generated using the same model — the standard coalescent for sample of size ten. Therefore, the variation among the trees reflects chance alone.

Effects of recombination. Recombination can be readily incorporated into the genealogical framework. The principle is the same as in traditional pedigree studies[9]. Recombination in a chromosomal segment means that it had two parental segments. So, the lineage of the segment splits in two, but precisely how the splitting occurs depends on the recombination process — for example, BACTERIAL CONJUGATION and meiotic crossing over have different effects on genealogies because, in the process of bacterial conjugation, the chromosome is necessarily broken in two places, whereas crossing over involves only a single break.

The main effect of recombination is that it allows linked sites to have different genealogical trees. To observe this, it is better to view recombination from a spatial rather than a temporal perspective. The genealogy of a sample of recombining sequences can be considered as a "walk through tree space"[10] — as we proceed from one end of the sequence to the other, the tree changes, but only gradually as each recombination event affects only a subset of the branches (FIG. 3). So, the extent to which the histories of different sites are correlated depends on the recombinational distance between them — as recombination approaches infinity, the genealogies of unlinked loci are conditionally independent, given the historical demography of the group under consideration. Because the pattern of polymorphism reflects the underlying genealogical trees, allele frequencies at linked sites in general cannot be independent. An important consequence of this dependency is linkage disequilibrium — the non-random association of alleles in haplotypes[11].

Recombination is very important to evolutionary inference, because unlinked or loosely linked loci can often be viewed as independent replicates of the evolutionary process. In the absence of recombination, the entire genome would correspond to a single genealogical tree, and we would never have more than a single independent replicate. So, the statistical benefits provided by recombination are substantial. As discussed below, the precision of evolutionary-inference methods increases rapidly with the number of genes studied, and very slowly with the number of sampled individuals.

What is the coalescent?

So far, we have used only the basic principles of Mendelian genetics to understand how genetic polymorphism data reflect the history of coalescence,

mutation and recombination. We now need a population-genetics model that incorporates these principles and that allows us to construct and analyse random genealogies. The coalescent has become the standard model for this purpose. This choice is not arbitrary, as the coalescent is a natural extension of classical population-genetics theory and models[7]. It was discovered independently by several authors in the early 1980s[12–15], although the definitive treatment is due to Kingman[12,16].

The basic idea underlying the coalescent is that, in the absence of selection, sampled lineages can be viewed as randomly 'picking' their parents, as we go back in time (FIG. 4). Whenever two lineages pick the same parent, their lineages coalesce. Eventually, all lineages coalesce into a single lineage, the MRCA of the sample. The rate at which lineages coalesce depends on how many lineages are picking their parents (the more lineages, the faster the rate) and on the size of the population (the more parents to choose from, the slower the rate).

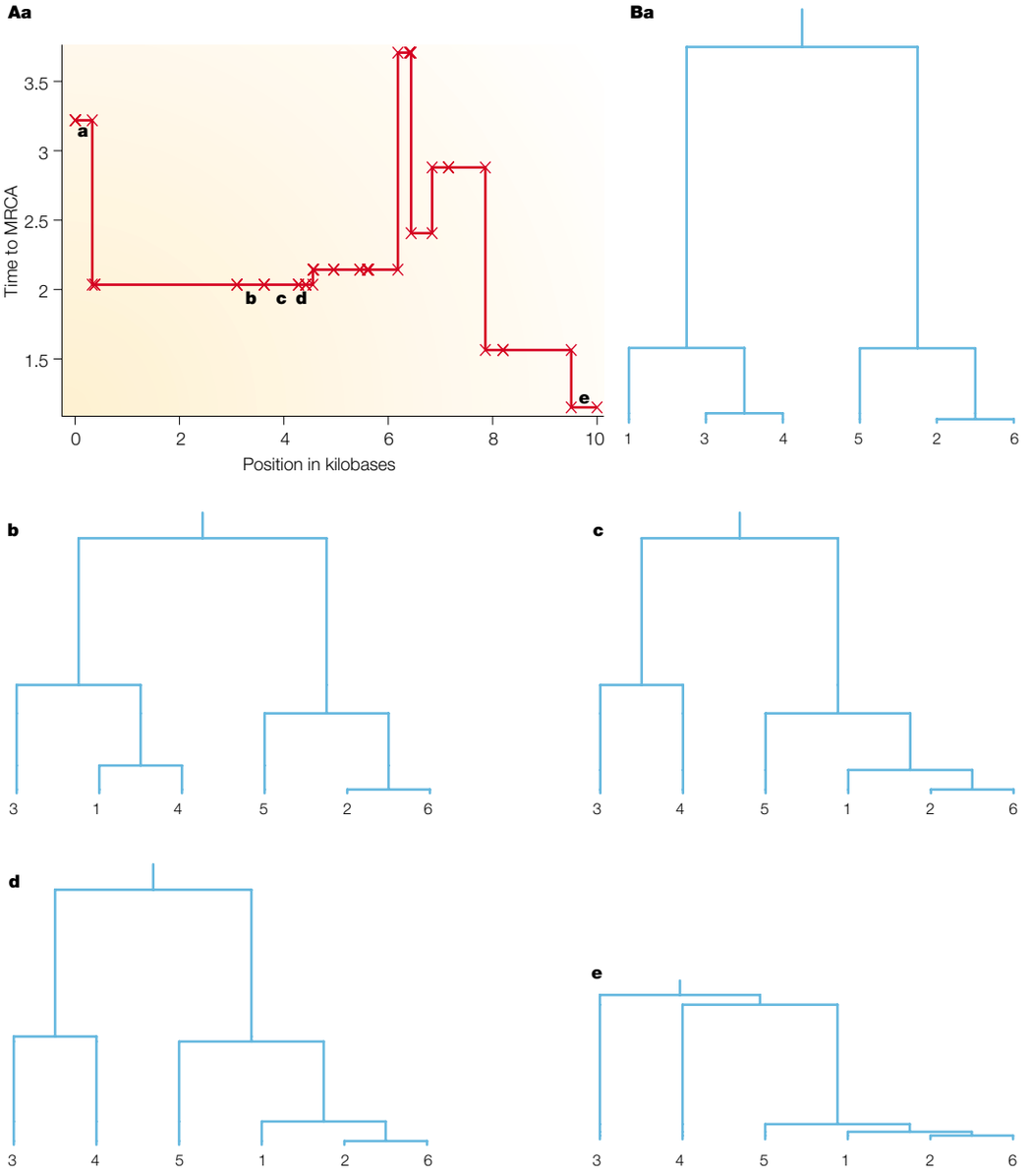


Figure 3 | **A simulated sample of six haplotypes using the standard coalescent with recombination. A** | In the top graph, the red line shows how the time to the most recent common ancestor (MRCA) (in units of coalescence time — 1 unit corresponds to 2*N* generations, if *N* is the size of the population) varies along the chromosome as a result of recombination. The parameters were chosen to represent ~10 kb of human DNA. The crosses along this line mark positions at which recombination took place in the history of the sample. Note that only a fraction of the recombination events resulted in a change of the time to the MRCA. **B** | A selection of gene trees (**a**–**e**) that correspond to specific positions along the chromosome (**a**–**e**) is shown. Trees for closely linked regions tend to be very similar (for example, **c** and **d**), if not identical. Numbers 1–6 represent individual haplotypes.

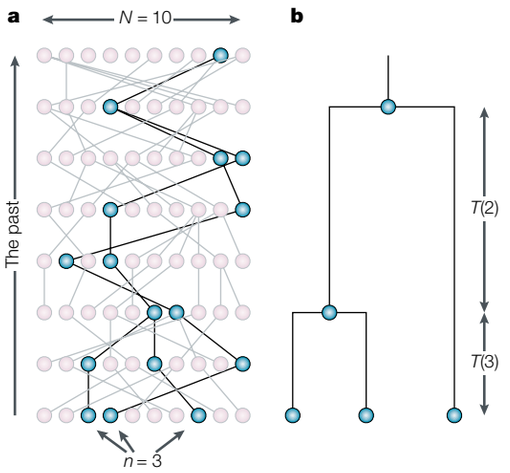


Figure 4 | **The basic principle behind the coalescent.** **a** | The complete genealogy for a population of ten haploid individuals is shown (diploid populations of *N* individuals are typically studied using a haploid model with 2*N* individuals[7]). The black lines trace the ancestries of three sampled lineages back to a single common ancestor. **b** | The subgenealogy for the three sampled lineages. In the basic version of the coalescent, it is only necessary to keep track of the times between coalescence events (*T*(3) and *T*(2)) and the topology — that is, which lineages coalesce with which. *N*, number of allelic copies in the population; *n*, sample size.

Because selectively neutral mutations do not affect reproduction, they can be superimposed on the tree afterwards.

Many other factors can be included in the model[7]. Some phenomena, such as variation in reproductive success, age structure and skewed sex ratios, change only the rate of coalescence, but other factors, such as population structure or fluctuation in population size, also change the shape of genealogical trees. Recombination has profound effects on the process, in that the coalescent with recombination does not generate a random tree, but rather a random graph[11,14,17]. This complication is, however, readily incorporated into the model. The only factor that causes real, although not insurmountable[18–20], difficulties is selection. By definition, under selection, some genotypes reproduce more than others, which means that, going back through time, lineages do not randomly pick parents.

The coalescent is closely related to classical population-genetics theory. The difference lies mainly in the type of question asked and the manner in which they are explored. Imagine that we wish to use a stochastic simulation to investigate the distribution of a sample statistic under a model that involves random GENETIC DRIFT. Traditionally, we would have simulated the evolution of the entire population, forwards in time, until equilibrium is reached (in other words, until the dependence on the starting conditions vanishes), and only then would a sample have been taken. The same procedure would have been repeated for each sample. Using the coalescent, we simulate the genealogy of the sample going back in time until the MRCA is found, and then add mutations forwards along the branches of the newly generated tree. Because we only use the individuals that are ancestral to the sample, there is no need to keep track of the entire population, and computational efficiency is greatly increased. The basic models, however, are the same; for studies of the effects of past evolutionary forces on current genetic variation, the coalescent is simply a better way to solve problems that could in principle (but not usually in practice) be solved using classical population genetics. Conversely, the classical forward-in-time approach is more appropriate for studies of how the long-term behaviour of evolutionary systems depends on initial conditions[7,21,22].

GENETIC DRIFT
The random fluctuations in allele frequencies over time that are due to chance alone.

HORIZONTAL TRANSFER
The transfer of genetic material between members of the same generation, or between members of different species.

ESTIMATOR
A function that produces an estimate of some parameter.

Why not phylogenetics?

When analysing polymorphism data, it is important to distinguish genealogical methods, such as those based on the coalescent, from methods that are borrowed from phylogenetics. Although both approaches involve trees, they are fundamentally very different.

Phylogenetic methods estimate trees. They were developed to determine the pattern of species descent, which is assumed to be tree-like. A single sequence from each species of interest is usually analysed, and the genealogy of the sequences is estimated. The estimated gene tree is then used to draw conclusions about relationships between species. Typically, the gene tree is simply equated with the species tree, an assumption that can be justified by the strong correlation between gene trees and species trees that is expected in most situations (BOX 1). However, the same approach makes little sense for questions that involve more complicated, demographic scenarios. In such cases, conclusions about the population tree cannot be drawn simply by looking at the estimated gene tree — different genes might produce different trees — and it is necessary to consider the likelihood of the estimated tree under alternative models. Furthermore, it might not make sense to try to estimate a population tree — the relevant model might involve migration (or HORIZONTAL TRANSFER) between populations, population history might not be tree-like, and the rates of migration might be of primary interest.

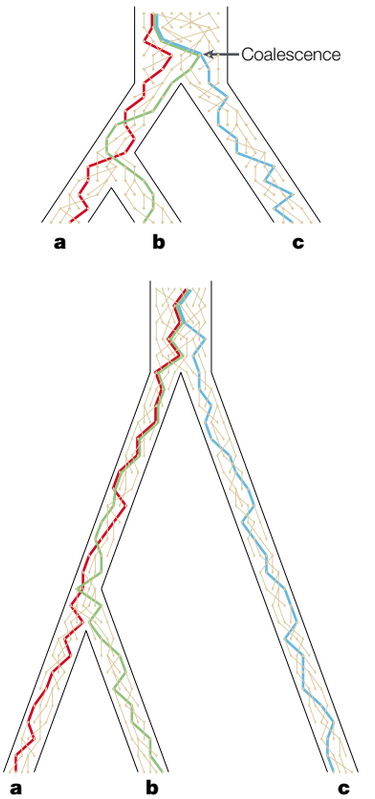
Genealogical methods do not estimate trees. Instead, they are used to estimate parameters of the random genealogical process that has given rise to each tree. In statistical terms, the tree itself typically becomes a nuisance parameter that is in itself of no inherent interest (BOX 2). The genealogical approach has none of the limitations of the phylogenetic methods and provides a coherent statistical framework in which to consider recombination, migration, selection and other processes.

Using the coalescent

As a tool for data analysis, the coalescent has many applications. Here, we consider its use as a mathematical modelling tool, as a simulation tool for hypothesis testing and for exploratory data analysis, and as the basis for full-likelihood inference.

Mathematical results. The coalescent process is a powerful mathematical tool that can be used to derive estimators of population parameters, such as rates of mutation or migration, and to devise statistical tests of models of evolution. For example, the widely used test that involves Tajima's *D* statistic[23] is based on the relationship between the average number of pairwise differences in a DNA sequence sample and the total number of observed mutations that is predicted by the basic coalescent model. If an unusual value of the *D* statistic is observed, the standard model might be rejected. It is often possible to show mathematically

Box 1 | **Gene trees and species trees**



The basic phylogenetic model relates species to each other through a bifurcating tree. This 'species tree' is estimated as the estimated genealogy of genes sampled from the different species. How can this approach be valid if, as we argue, each gene tree is the random outcome of a historical process? Different genes should give rise to different trees; in fact, even a single gene could have many trees as a result of intragenic recombination. This apparent paradox is resolved by the fact that, as long as time intervals between species-branching events are much greater than time intervals between lineage-branching events in each species, gene and species divergences are likely to be nearly concurrent[64,78], and gene trees are likely to be very similar to the species tree[13,71,72,79].

Consider the two species trees that are shown here. Embedded in each species tree is the gene tree that relates all existing copies of a particular gene to each other. The gene trees are random: each copy of a gene is connected to a random copy from the previous generation. As illustrated by the top tree, it is possible for a gene tree not to reflect the species tree. A gene sampled from species **b** (green) and one from species **c** (blue) are more closely related to each other than to a gene sampled from species **a**, even though **a** and **b** were the last species to separate. By contrast, in the bottom tree, all lineages in a species coalesce more recently with each other than do the species themselves. Therefore, it matters neither how many copies are sampled from each species, nor which copies are sampled — all samples indicate the same tree shape. Furthermore, trees for different genes (as well as for recombinant parts of the same gene) will all have this same topology. The lengths of gene-tree branches might differ (depending on how long it takes to reach the most recent common ancestor (MRCA) in the ancestral species), but these random variations are small compared with the total length of each branch. The difference between the examples shown here is that, in the top tree, the branches of the species tree are of similar length to the branches of the genealogical trees in each species, whereas, in the bottom tree, the branches of the species tree are much longer. How long must the species-tree branches be to avoid discordance of gene trees? The answer depends on the within-species evolutionary model. Under the basic coalescent model, the expected time to the MRCA for the entire population is twice the 'effective' number of copies of the gene in the population, in units of generations (in humans, estimates of the time to the MRCA are typically of the order of one million years[36]). So, branches of the species tree that are several times longer than average times to within-species MRCAs should usually suffice to avoid discordance. However, because genealogical trees are random, there is always a small probability that a particular gene tree will disagree with the species tree (this probability will be higher for loci with genealogies that have been deepened by BALANCING SELECTION[80,81]).

Because large phylogenies are more likely to contain some short species branches, we expect them to produce some discordance. We also expect discordance between gene trees and species trees for phylogenies that include ancient but rapid divergence, such as ADAPTIVE RADIATIONS of African cichlid fish[82]. Nonetheless, it is safe to say that the phylogeny of well-separated species, such as human, cow and chicken, is not affected by the randomness of genealogies.

For more closely related species, however, gene trees and species trees often disagree[13,83–85]. Such discordance can be caused, for example, by gene exchanges[72,86], but random genealogies often provide a simple explanation for this observation. In these cases, it is useful to infer the species tree using more than one individual per species[45,84,87] and, more importantly, using more than one locus[29,88,89]. For example, Chen and Li[90] obtained gene trees for 53 randomly chosen, non-coding regions in human, gorilla and chimpanzee. Of these, 31 supported human–chimp, 10 supported human–gorilla, and 12 supported chimp–gorilla as the most closely related pair, and a comparison of the likelihoods of all three models shows that the human–chimp grouping is statistically supported with near certainty.

It should be noted that genes that had direct roles in the divergence of incipient species are likely to have diverged simultaneously with species, so that their genealogies will more accurately reflect species trees[71,91,92]. For example, it is thought that selection on a region of the *tb1* (*teosinte branched1*) gene was involved in the divergence of maize from teosinte. Phylogenies that are based on this region perfectly separate maize and teosinte sequences, whereas those based on other regions do not[93]. A similar result holds for the hybrid sterility gene *OdsH* (*Ods-site homeobox*) in *Drosophila melanogaster*[92].

BALANCING SELECTION
The selection that maintains two or more alleles in a population.

ADAPTIVE RADIATION
The evolution of new species or subspecies to fill unoccupied ecological niches.

Box 2 | **Likelihood for trees**

Basic statistics makes the distinction between phylogenetic and coalescent approaches apparent. The fundamental equation for likelihood inference in phylogenetics is[94]

$$L = P(D|G, \boldsymbol{\mu}), \quad (1)$$

where L is the likelihood (the probability of the data, given the parameters), D is the data (typically DNA sequences), G is the tree and μ is the collection of parameters in the mutation model. The objective of the analysis is to estimate the parameter G.

The analogous equation in the coalescent setting is[22,42,94]

$$L = \sum_{G} P(D|G, \boldsymbol{\mu})P(G, \boldsymbol{\alpha}), \quad (2)$$

where α is the collection of parameters (such as population sizes and migration rates) for the population process. The objective of the analysis is typically to estimate these parameters. The tree or genealogy, G, is a so-called nuisance parameter, which we remove by averaging the likelihood over all possible values.

In the event that features of G (or G itself) are of interest, it is more natural to treat them as random variables than as parameters to be estimated[5]. Because there is only one actual evolutionary history, it might be argued that a BAYESIAN statistical viewpoint, as opposed to a FREQUENTIST APPROACH, is warranted[44]. From a Bayesian perspective, the coalescent provides the most accurate characterization of genealogies that can be made before data are observed.

how departures from the standard model (such as those caused by population structure or selection[7]) affect the TEST STATISTIC, which makes it possible to interpret the observed deviation.

Coalescent theory also provides insights into the peculiarities of population-genetics data. A good example is the effect that sample size has on data analysis. Because all existing copies of a particular sequence in the genome are related through a genealogical tree, population samples are never completely independent[24]. The three copies in FIG. 1 that have the derived base T rather than the ancestral G did not mutate independently from G to T. Instead, they share T because of common ancestry. This obvious fact has surprising statistical consequences. Most importantly, increasing the sample size, n, does not improve the accuracy of estimates in the manner we are used to in conventional statistical analysis. For example, in the standard coalescent, the VARIANCE of estimators of the scaled mutation rate $\theta = 4Nu$ decreases at a rate of $1/\log n$, rather than $1/n$. This means that increasing the sample size is only marginally effective in improving the estimate. The reason for this is that no matter how large the sample, there is still only a single underlying genealogical tree.

The degree to which sampled sequences are correlated depends on the evolutionary model — for example, in a rapidly growing population, sequences are less correlated[24]. The only way that variance caused by the random nature of the trees can be reduced is by considering recombination — that is, by observing several loci.

Another illuminating example concerns the MRCA of a population. One might imagine that quite a large sample is required to ensure that the MRCA of the sample is also the MRCA of the entire population, but it can be shown[25] that, for an unstructured population, this probability is simply $(n-1)/(n+1)$. So, even the genealogy for a small sample is likely to contain the MRCA of a population.

A wide variety of evolutionary scenarios can be modelled using the coalescent. Even if the mathematical analysis is intractable, it is almost always easy to simulate the process. This makes it possible to investigate whether a particular scheme can be expected to have left a trace in the data, which is useful to know before embarking on an empirical study to investigate the scenario in question. Historical events, such as BOTTLENECKS and migrations, can be surprisingly difficult to detect. Many questions in human evolution are plainly unanswerable using data that have so far been available[26,27] (BOX 3). The coalescent can provide useful guidance about how many individuals, populations and loci need to be sampled to answer the questions of interest[27–29].

Coalescent simulations. One of the most widespread uses of the coalescent is as a simulation tool[4,7]. Using the coalescent, it is possible to simulate samples from a wide variety of models. Compared with the alternative, classical population-genetics simulations — which run forwards in time — coalescent simulations are easier to implement and much more efficient.

Coalescent simulations are very well suited to exploratory data analysis. Samples that are simulated under various models can be combined with data to test hypotheses. The canonical approach, developed mainly by Hudson and colleagues[4,30,31], can be described as follows. Let us assume that we observe an unusual pattern of polymorphism in a data set and want to know if it is simply a historical accident or whether it requires a special explanation (such as the existence of selection). We then simulate many possible data sets under a null coalescent model that does not include the factor of interest (for example, selection). Finally, we compare the value of a test statistic obtained from the real data with the distribution of values obtained from the simulated data sets. If patterns that are characteristic of the actual data are rarely seen in the simulations, we reject the null hypothesis. Having rejected the null hypothesis, it is usually possible to propose patterns that are compatible with the actual data. It is typical of population-genetics applications that many alternative hypotheses are possible. It is therefore important to choose carefully the alternatives for characterizing deviations from the null. It should also be kept in mind that the procedure outlined above is often a form of post-hoc analysis that suffers from the usual problems of multiple comparisons.

A prototypical coalescent simulation study is given by Hudson *et al.*[32]. At the *Superoxide dismutase* (*Sod*) locus in *Drosophila melanogaster*, Hudson *et al.* studied DNA sequence variation in a 1,410-bp region. Surprisingly, they found that five out of ten sequences in a sample from Barcelona, Spain, were identical in this region, sharing an allele termed *Fast-A*. The remaining five sequences were all different and contained 55 polymorphic sites. To test if this type of data

BAYESIAN APPROACH
A statistical perspective that focuses on the probability distribution of parameters, before and after seeing the data.

FREQUENTIST APPROACH
A statistical perspective that focuses on the frequency with which an observed value is expected in numerous trials.

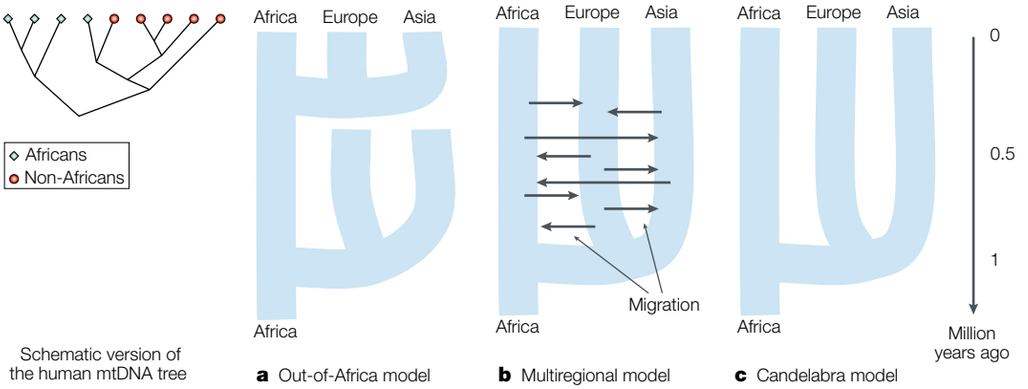
TEST STATISTIC
A function that produces values from data for comparing with expected values under various models.

VARIANCE
A statistic that quantifies the dispersion of data about the mean.

BOTTLENECK
A temporary marked reduction in population size.

Box 3 | **The meaning of 'mitochondrial Eve'**

The difference between phylogenetic and coalescent approaches is well illustrated by the interpretation of the human mitochondrial DNA (mtDNA) tree. The first estimates of this tree caught the world's attention by indicating that all modern humans shared a common female ancestor as recently as 200,000 years ago, and that the root of the tree was among Africans[95,96]. The tree was interpreted as evidence for the 'out-of-Africa' model of human origins, in which modern humans evolved in Africa and spread over the world relatively recently (perhaps 100,000 years ago), replacing *Homo erectus*, which had dispersed worldwide much earlier. The studies were immediately criticized on several methodological grounds, the most important being that the estimated tree was by no means the only possible one, as several equally plausible trees had a non-African root[97–99]. Subsequent studies have tended to support the original tree[100], but a more fundamental problem is that the mtDNA tree itself actually tells us very little about human origins. The problem is simply that the tree is compatible with many reasonable hypotheses. The figures below show cartoon versions of the data, and three popular models[101,102]. Because the root is too recent, the mtDNA tree is not compatible with the 'candelabra' model, which can therefore be rejected. However, the tree could have arisen under either the out-of-Africa or the 'multiregional' model of human origins. Under the out-of-Africa model, a recent African root is expected; under the multiregional model, an African root is one of three possibilities, and a recent root is common in models that include migration. Therefore, even if it were possible to estimate unambiguously the mtDNA tree, we would be unable to choose between these two models, unless we were able to compare the likelihood that each of them could have given rise to the observed tree. This cannot be done using phylogenetic methods, but is precisely what coalescent methods are designed to do (BOX 2). With respect to the question of human origins, an immediate conclusion is that a single gene tree will almost never suffice to choose among demographic models of population histories[26,27].



set was unusual under a model of neutral evolution acting on that locus, the authors simulated 10,000 samples. In each, they simulated random genealogies under the standard coalescent model, both with and without recombination, using sample sizes of ten. They then randomly placed 55 mutations on the gene tree, obtaining random data sets (it should be noted that to base the analysis on the observed number of mutations rather than on the unknown mutation rate is not correct technically, but in practice, the problem is only an issue for very small values of the mutation parameter[33–35]). For each data set that was constructed in this way, they checked whether it included a subset of five haplotypes that had no polymorphism (FIG. 5). For the model with no recombination, only 1.1% of the runs included such a subset, leading Hudson *et al.* to suggest an alternative hypothesis of favourable selection or 'hitchhiking' acting on the *Fast-A* variant.

A feature of the simulation approach is that the null hypothesis does not have to be a basic coalescent model. For example, Takahata *et al.*[36] used the simulation approach to investigate whether a particular multiregional model of human origins was compatible with data from ten human loci. Using a model of migration between three subpopulations — African, European and Asian — followed by their divergence (BOX 3), they simulated many genealogies and for each, they determined the position of the MRCA. In the empirical data, nine out of ten estimated genealogies had an African ancestor (and the tenth locus had only a single polymorphic site). This pattern was found to be highly unlikely under the investigated multiregional model, unless the African population size was much larger than the Asian and the European. As data accumulate for more loci, it should be possible to test increasingly sophisticated models of human evolution in this manner.

Coalescent simulations can also be used for purposes that are only indirectly related to data analysis. As mentioned above, they can be useful in study design — for example, to determine the number of loci that need to be surveyed. Additionally, simulated samples help to evaluate the performance of new statistical tests[37,38]. This approach can be valuable whether or not the proposed tests are based on coalescent reasoning[39]. Because some methods are developed in advance of the appropriate data to which they can be applied, coalescent simulations conveniently provide data sets on which new methods can be tested. Finally, as discussed below, coalescent simulations can have an important role in likelihood calculations.

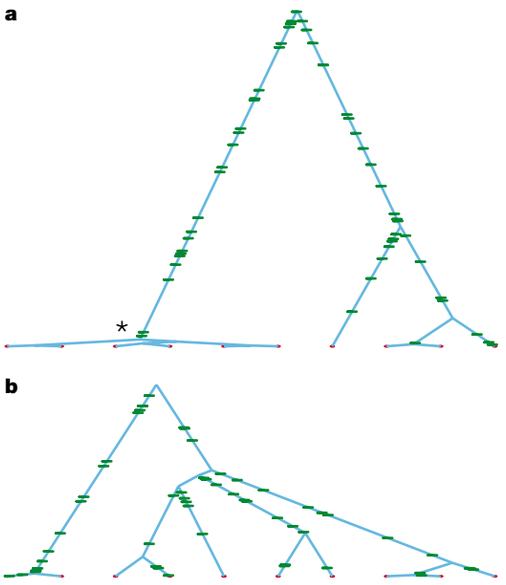


Figure 5 | **Simulated random genealogies with a sample size of ten.** The figure is based on data from the study of Hudson *et al.*[32]. Both genealogies were simulated from the standard coalescent model (without recombination). A total of 55 mutations (green horizontal lines) were placed randomly on each tree. **a** | One set of at least five sequences with no polymorphism; (marked by an asterisk) **b** | All sets of sequences contained polymorphisms. Red dots denote individual haplotypes.

Likelihood methods. From a statistical point of view, one of the most exciting aspects of the coalescent is that it allows full LIKELIHOOD ANALYSIS of evolutionary models[8]. The idea is straightforward in theory. All we need to do is evaluate the likelihood EQN (2) in BOX 2 for our data and for our favourite models. Unfortunately, this is not easy in practice, because summing (or integrating) over all possible genealogies turns out to be exceedingly difficult. Advanced computational techniques, such as IMPORTANCE SAMPLING[40,41] and MARKOV CHAIN MONTE CARLO[42–44], have been applied to this problem, but it has so far been difficult to obtain reliable results, except for very simple models. Encouragingly, some progress with more sophisticated demographic models that include population divergence[45–47], migration[48,49] or both[50,51] has recently been made.

A promising alternative is to use approximate methods based on SUMMARY STATISTICS[52–55]. Instead of numerically evaluating EQN (2) (BOX 2), the full data set *D* is replaced by a set of summary statistics, *S*, such as the number of observed SEGREGATING SITES. Data are then simulated using various parameter values, and each simulation is accepted or rejected according to how close the simulated data are to the real data. For a given set of parameter values, the likelihood is then approximated by the proportion of simulations that are accepted.

This summary-statistic approach has great potential for the inference of parameters under models for which complete evaluation of EQN (2) is intractable. Important for its success is the extent to which values of summary statistics capture the information in full data sets. Careful choices of such statistics will be needed for likelihood computations in complex models.

Software. The coalescent is a modelling tool that can be used in a wide variety of situations. Consequently, the diversity of conceivable coalescent-based analyses is too vast for any single software package to encompass. Several standard estimators and tests can be computed using packages such as DnaSP[56] and SITES[57]. Researchers might also construct programs that suit their particular needs, using readily available subroutines for coalescent simulations[58].

The importance of pre-existing software packages is greatest for the third category of coalescent-based analysis — the likelihood methods. A summary of computer programs that are useful in likelihood analysis, including LAMARC[43,59], GENETREE[40] and BATWING[47], is given in REF. 8.

Conclusion

The analysis of polymorphism data must take the historical nature of the data into account. Today, polymorphism data are often analysed using methods that are borrowed from phylogenetics, in an approach that can be represented in the following way. First, collect sequence data. Second, estimate the genealogical tree of the sample sequence (without regard for recombination, that is, regardless of whether or not such a tree really exists). Finally, tell a story based on the estimated tree. The approach favoured by us can be summarized as follows (see also BOX 2). First, collect sequence data; second, consider all possible genealogies, including those with recombination, and their probabilities under models of interest. Third, for each genealogy, calculate the likelihood of the data (the total likelihood of the data under each of the models is the sum of the likelihoods for all trees, weighted by the probability of each genealogy under that model). Finally, estimate parameters by finding values that maximize the likelihood of the data, and test models by comparing likelihoods under different hypotheses.

The advantages of the second approach include its ability to choose among models using standard statistical criteria, and its capacity for incorporating migration, ADMIXTURE and other demographic phenomena. Perhaps most importantly, in the era of genomic polymorphism data, the coalescent approach can naturally incorporate the effects of recombination, whereas the phylogenetic approach cannot.

The use of phylogenetic methods is justified if interest lies not in the parameters of the evolutionary model, but rather in the particular history of a specific locus. This is sometimes true, for example, when Y chromosomes are used to study patrilineal inheritance of surnames[60]. Even in these situations, however, relying on an estimated genealogy before proceeding with the analysis ignores uncertainty in the branching times, if not also the branching order itself. Typically, the analysis leads to artificially reduced confidence intervals, because this source of error is not taken into account, as it is in the coalescent approach.

LIKELIHOOD ANALYSIS
A statistical method that considers the likelihood of observing the data under alternative models.

IMPORTANCE SAMPLING
A computational technique for efficient numerical calculation of likelihoods.

MARKOV CHAIN MONTE CARLO
A computational technique for efficient numerical calculation of likelihoods.

SUMMARY STATISTIC
A function that summarizes complex data in terms of simple numbers (examples include mean and variance).

SEGREGATING SITE
A DNA base-pair position at which polymorphism is observed in a population.

ADMIXTURE
The mixing of two genetically differentiated populations.

More importantly, in the usual analyses of single loci, a particular realization of a random genealogical process is not of great interest, and there is a danger of drawing unwarranted general conclusions. To take the familiar example of human origins (BOX 3), we are surely interested in the history of human migrations, not in the genealogical relationships between mitochondria or Y chromosomes *per se*.

Problems with the coalescent. It is one thing to say that random genealogies should be taken into account when studying polymorphisms, and quite another to say how this should be done. The coalescent is a natural choice for modelling genealogies, and there are strong theoretical reasons to believe that it is often a good model for actual data[7]. However, the fact remains that it makes an uncomfortable number of assumptions (for example, about the absence of selection[61]). It seems to us that the reliance on specific models has largely been a consequence of the scarcity of data. As genomic polymorphism data become common, it should be possible to rely more and more on empirically estimated, coalescent-based models. For example, most tests of selection are based on the rejection of a coalescent model that assumes a population without geographical structure[30]. This is problematic because geographical structure can affect patterns of polymorphism in ways that mimic selection[7], which increases the risk of false positives. This problem could be overcome by first estimating the population structure using genomic data, and then by testing for selection using the estimated population-structure model to construct the null hypothesis. Techniques for carrying out this kind of test have been developed in the context of linkage-disequilibrium (LD) mapping[39,62], which is also troubled by false positives that are caused by population structure.

A second set of potential problems concerns the computational difficulties of some coalescent-based methods. Recombination, in particular, makes likelihood-based inference exceedingly difficult. For many questions that researchers might naturally wish to address using polymorphism data, suitable coalescent-based methods have not yet been devised. It remains a challenge for theorists to anticipate forthcoming data and to develop appropriate analysis techniques. To analyse data on a genomic scale, approximate methods will almost certainly be necessary.

Prospects. Coalescent theory has revolutionized molecular population genetics over the past 20 years. It seems certain that it will have a similarly marked impact on fields such as molecular ecology, PHYLOGEOGRAPHY and the study of human origins, as indicated by some of its recent uses[36,54,63,64]. One other area in which the coalescent might be of use is the evolutionary genetics and epidemiology of infectious disease[65,66]. Important questions about epidemics of infectious bacterial and viral agents concern the demographic parameters of pathogen populations, and therefore fall in the realm of coalescent methods. Examples that seem appropriate for coalescent-based analysis include the timing of introduction of a pathogen into the human population, whether certain virulent lineages have been favoured by selection and how rapidly the evolution of pathogens occurs in hosts. Serially sampled data that are typical of host–pathogen systems raise new theoretical issues, and this area will probably become an important source of problems for coalescent methods[67,68].

A genealogical approach might also affect how the origin and divergence of species are studied. Variation among gene genealogies across loci means that a view of evolution as the bifurcation of complete genomes is not always tenable[69,70]. As described in BOX 1, it is sufficiently common for genes to differ in their histories that such discordance cannot be treated as unusual. Sexual organisms can be viewed as collections of genomic regions with different histories[71,72]; for asexual organisms, bacterial recombination (that is, horizontal transfer) might be sufficiently widespread that a similar perspective must be adopted.

Coalescent-based methods can be used for inference of relationships among groups that are sufficiently related that genealogies of different genomic regions disagree. Methods of estimating population and species trees must be adapted to accommodate equally valid but differing histories presented by different parts of the genome, as well as polymorphisms shared among closely related species. Coalescent-based methods for tree estimation, which, unlike most traditional phylogenetic algorithms, are designed to allow variation both in and between species, are uniquely suited for doing this[45,46].

Finally, coalescent-based methods have the potential to assist in LD mapping of genes that underlie complex traits[11]. LD mapping uses the pattern of associations among genotypes and phenotypes that arises from the history of recombination, coalescence and mutation to identify disease-susceptibility loci. The objective is not to test evolutionary models, but to use the pattern of LD that actually exists. Coalescent-based simulations have been used to explore the properties of LD in the human genome[73,74]. Additionally, reasoning based on simple coalescent models underlies new methods for identifying disease-susceptibility genes[75,76]. As the actual pattern of LD in the human genome becomes clearer[77], more realistic genealogical models can be applied.

In summary, we are convinced that it is essential to appreciate the dependence of genetic variation on its underlying genealogies to analyse polymorphism data in a rigorous statistical framework. The coalescent provides a method to model this dependence. As genomic data proliferate, its importance is only likely to increase.

PHYLOGEOGRAPHY
The use of estimated gene genealogies to study geographical history and structure of populations and species.

1. Luria, S. E. & Delbrück, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**, 491–511 (1943).
2. Chakravarti, A. Population genetics — making sense out of sequence. *Nature Genet.* **21**, S56–S60 (1999).
3. Tavaré, S. Line-of-descent and genealogical processes, and their applications in population genetic models. *Theor. Popul. Biol.* **26**, 119–164 (1984).
4. Hudson, R. R. in *Oxford Surveys in Evolutionary Biology* Vol. 7 (eds Futuyma, D. & Antonovics, J.) 1–43 (Oxford Univ. Press, Oxford, UK, 1990).
5. Donnelly, P. & Tavaré, S. Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* **29**, 401–421 (1995).
6. Fu, Y.-X. & Li, W.-H. Coalescing into the 21st century: an overview and prospects of coalescent theory. *Theor. Popul. Biol.* **56**, 1–10 (1999).
7. Nordborg, M. in *Handbook of Statistical Genetics* (eds Balding, D. J., Bishop, M. J. & Cannings, C.) 179–212 (John Wiley & Sons, Chichester, UK, 2001).
8. Stephens, M. in *Handbook of Statistical Genetics* (eds Balding, D. J., Bishop, M. J. & Cannings, C.) 213–238 (John Wiley & Sons, Chichester, UK, 2001).
References 7 and 8 provide current technical reviews of the coalescent and its use in evolutionary inference.
9. Thompson, E. A. *Statistical Inference from Genetic Data on Pedigrees* (Institute of Mathematical Statistics, Beachwood, Ohio, 2000).
10. Wiuf, C. & Hein, J. Recombination as a point process along sequences. *Theor. Popul. Biol.* **55**, 248–259 (1999).
11. Nordborg, M. & Tavaré, S. Linkage disequilibrium: what history has to tell us. *Trends Genet.* **18**, 83–90 (2002).
12. Kingman, J. F. C. On the geneaology of large populations. *J. Appl. Prob.* **19A**, 27–43 (1982).
This paper provides the first description of the coalescent.
13. Hudson, R. R. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* **37**, 203–217 (1983).
14. Hudson, R. R. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**, 183–201 (1983).
15. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
16. Kingman, J. F. C. Origins of the coalescent: 1974–1982. *Genetics* **156**, 1461–1463 (2000).
17. Griffiths, R. C. & Marjoram, P. Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3**, 479–502 (1996).
18. Kaplan, N. L., Darden, T. & Hudson, R. R. The coalescent process in models with selection. *Genetics* **120**, 819–829 (1988).
19. Neuhauser, C. & Krone, S. M. The genealogy of samples in models with selection. *Genetics* **145**, 519–534 (1997).
20. Slatkin, M. Simulating genealogies of selected alleles in a population of variable size. *Genet. Res.* **78**, 49–57 (2001).
21. Ewens, W. J. in *Mathematical and Statistical Developments of Evolutionary Theory* (ed. Lessard, S.) 177–227 (Kluwer Academic, Dordrecht, 1990).
22. Felsenstein, J. in *Evolutionary Genetics: From Molecules to Morphology* Vol. 1 Ch. 29 (eds Singh, R. S. & Krimbas, C. B.) 609–627 (Cambridge Univ. Press, New York, 2000).
A readable and amusing overview of the history of population genetics.
23. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
24. Donnelly, P. in *Variation in the Human Genome* 25–50 (Ciba Foundation–Wiley, Chichester, UK, 1996).
This paper lucidly describes the importance of incorporating genealogy in studies of genetic polymorphism.
25. Saunders, I. W., Tavaré, S. & Watterson, G. A. On the genealogy of nested subsamples from a haploid population. *Adv. Appl. Prob.* **16**, 471–491 (1984).
26. Nordborg, M. On the probability of Neanderthal ancestry. *Am. J. Hum. Genet.* **63**, 1237–1240 (1998).
27. Wall, J. D. Detecting ancient admixture in humans using sequence polymorphism data. *Genetics* **154**, 1271–1279 (2000).
28. Pluzhnikov, A. & Donnelly, P. Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* **144**, 1247–1262 (1996).
The authors describe the effect of recombination in reducing the variation of estimates of evolutionary parameters.
29. Wu, C.-I. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. *Genetics* **127**, 429–435 (1991).
30. Kreitman, M. Methods to detect selection in populations with applications to the human. *Annu. Rev. Genomics Hum. Genet.* **1**, 539–559 (2000).
31. Nielsen, R. Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**, 641–647 (2001).
References 30 and 31 describe how the signature of selection in DNA sequence polymorphism might be detected.
32. Hudson, R. R., Bailey, K., Skarecky, D., Kwiatowski, J. & Ayala, F. J. Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. *Genetics* **136**, 1329–1340 (1994).
33. Markovtsova, L., Marjoram, P. & Tavaré, S. On a test of Depaulis and Veuille. *Mol. Biol. Evol.* **18**, 1132–1133 (2001).
34. Wall, J. D. & Hudson, R. R. Coalescent simulations and statistical tests of neutrality. *Mol. Biol. Evol.* **18**, 1134–1135 (2001).
35. Depaulis, F., Mousset, S. & Veuille, M. Haplotype tests using coalescent simulations conditional on the number of segregating sites. *Mol. Biol. Evol.* **18**, 1136–1138 (2001).
36. Takahata, N., Lee, S.-H. & Satta, Y. Testing multiregionality of modern human origins. *Mol. Biol. Evol.* **18**, 172–183 (2001).
37. Wakeley, J. Distinguishing migration from isolation using the variance of pairwise differences. *Theor. Popul. Biol.* **49**, 369–386 (1996).
38. Wall, J. D. Recombination and the power of statistical tests of neutrality. *Genet. Res.* **74**, 65–79 (1999).
39. Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
40. Griffiths, R. C. & Tavaré, S. Ancestral inference in population genetics. *Stat. Sci.* **9**, 307–319 (1994).
41. Stephens, M. & Donnelly, P. Inference in molecular population genetics. *J. R. Stat. Soc. B* **62**, 605–655 (2000).
42. 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).
43. Kuhner, M. K., Yamato, J. & Felsenstein, J. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149**, 429–434 (1998).
44. Wilson, I. J. & Balding, D. J. Genealogical inference from microsatellite data. *Genetics* **150**, 499–510 (1998).
45. Nielsen, R. Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. *Theor. Popul. Biol.* **53**, 143–151 (1998).
46. Nielsen, R., Mountain, J. L., Huelsenbeck, J. P. & Slatkin, M. Maximum likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution* **52**, 669–677 (1998).
47. Wilson, I. J., Weale, M. E. & Balding, D. J. Inferences from DNA data: population histories, evolutionary processes, and forensic match probabilities. *J. R. Stat. Soc. A* (in the press).
This is a good example of the likelihood framework. Likelihoods of hierarchical divergence schemes are compared. Using Y-chromosome data, the model supports a division between African and non-African populations for the most ancient human divergence.
48. Bahlo, M. & Griffiths, R. C. Inference from gene trees in a subdivided population. *Theor. Popul. Biol.* **57**, 79–95 (2000).
49. 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).
50. Nielsen, R. & Slatkin, M. Likelihood analysis of ongoing gene flow and historical association. *Evolution* **54**, 44–50 (2000).
51. Nielsen, R. & Wakeley, J. Distinguishing migration from isolation: a Markov Chain Monte Carlo approach. *Genetics* **158**, 885–896 (2001).
This paper shows considerable progress on a problem that has been notoriously difficult to solve with such methods as genetic-distance analysis, namely, distinguishing between ancient divergence followed by recent migration and recent divergence with no subsequent migration.
52. Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518 (1997).
A seminal paper that contains one of the first uses of summary statistics for approximate likelihood calculations, an approach which is likely to become increasingly important.
53. Weiss, G. & von Haeseler, A. Inference of population history using a likelihood approach. *Genetics* **149**, 1539–1546 (1998).
54. 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).
55. Wall, J. D. A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.* **17**, 156–163 (2000).
56. Rozas, J. & Rozas, R. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**, 174–175 (1999).
57. Hey, J. & Wakeley, J. A coalescent estimator of the population recombination rate. *Genetics* **145**, 833–846 (1997).
58. Hudson, R. R. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).
59. Beerli, P. & Felsenstein, J. Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proc. Natl Acad. Sci. USA* **98**, 4563–4568 (2001).
60. Jobling, M. A. In the name of the father: surnames and genetics. *Trends Genet.* **17**, 353–357 (2001).
61. Gillespie, J. H. Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics* **155**, 909–919 (2000).
62. Pritchard, J. K. & Donnelly, P. Case–control studies of association in structured or admixed populations. *Theor. Popul. Biol.* **60**, 227–237 (2001).
63. Ford, M. J. Testing models of migration and isolation among populations of chinook salmon (*Oncorhynchus tschawytscha*). *Evolution* **52**, 539–557 (1998).
64. Edwards, S. V. & Beerli, P. Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* **54**, 1839–1854 (2000).
65. Crandall, K. A. (ed.) *The Evolution of HIV* (Johns Hopkins Univ. Press, Baltimore, Maryland, 1999).
66. Thompson, R. C. A. (ed.) *Molecular Epidemiology of Infectious Diseases* (Arnold, London, 2000).
67. Rodrigo, A. G. *et al.* Coalescent estimates of HIV-1 generation time *in vivo*. *Proc. Natl Acad. Sci. USA* **96**, 2187–2191 (1999).
68. Fu, Y.-X. Estimating mutation rate and generation time from longitudinal samples of DNA sequences. *Mol. Biol. Evol.* **18**, 620–626 (2001).
69. Wu, C.-I. The genic view of the process of speciation. *J. Evol. Biol.* **14**, 851–865 (2001).
70. Rieseberg, L. H. & Burke, J. M. A genic view of species integration. *J. Evol. Biol.* **14**, 883–886 (2001).
71. Hey, J. in *Molecular Ecology and Evolution: Approaches and Applications* (eds Schierwater, B., Streit, B., Wagner, G. P. & DeSalle, R.) 435–449 (Birkhäuser, Basel, Switzerland, 1994).
72. Maddison, W. P. Gene trees in species trees. *Syst. Biol.* **46**, 523–536 (1997).
73. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
74. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
75. Liu, J. S., Sabatti, C., Teng, J., Keats, B. J. B. & Risch, N. Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res.* **11**, 1716–1724 (2001).
76. Morris, A. P., Whittaker, J. C. & Balding, D. J. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am. J. Hum. Genet.* **70**, 686–707 (2002).
References 75 and 76 show how the coalescent might be used for fine mapping of disease-susceptibility sites in a case–control setting.
77. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
78. Rosenberg, N. A. & Feldman, M. W. in *Modern Developments in Theoretical Population Genetics* ch. 9 (eds Slatkin, M. & Veuille, M.) 130–164 (Oxford Univ. Press, Oxford, UK, 2002).
79. Nichols, R. Gene trees and species trees are not the same. *Trends Ecol. Evol.* **16**, 358–364 (2001).
80. Takahata, N. & Nei, M. Allelic genealogy under overdominant and frequency dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* **124**, 967–978 (1990).
81. Ioerger, T. R., Clark, A. G. & Kao, T.-H. Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. *Proc. Natl Acad. Sci. USA* **87**, 9732–9735 (1990).
82. Takahashi, K., Terai, Y., Nishida, M. & Okada, N. Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in Lake Tanganyika as revealed by the insertion of retroposons. *Mol. Biol. Evol.* **18**, 2057–2066 (2001).
83. Pamilo, P. & Nei, M. Relationships between gene trees and species trees. *Mol. Biol. Evol.* **5**, 568–583 (1988).
84. Takahata, N. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* **122**, 957–966 (1989).

85. Wakeley, J. The effects of subdivision on the genetic divergence of populations and species. *Evolution* **54**, 1092–1101 (2000).
86. Eisen, J. A. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Curr. Opin. Genet. Dev.* **10**, 606–611 (2000).
87. Rosenberg, N. A. The probability of topological concordance of gene trees and species trees. *Theor. Popul. Biol.* (in the press).
88. Saitou, N. & Nei, M. The number of nucleotides required to determine the branching order of three species, with special reference to the human–chimpanzee–gorilla divergence. *J. Mol. Evol.* **24**, 189–204 (1986).
89. Ruvolo, M. Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. *Mol. Biol. Evol.* **14**, 248–265 (1997).
90. Chen, F.-C. & Li, W.-H. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68**, 444–456 (2001).
91. Palopoli, M. F., Davis, A. W. & Wu, C.-I. Discord between the phylogenies inferred from molecular versus functional data: uneven rates of functional evolution or low levels of gene flow? *Genetics* **144**, 1321–1328 (1996).
92. Ting, C.-T., Tsaur, S.-C. & Wu, C.-I. The phylogeny of closely related species as revealed by the genealogy of a speciation gene, *Odysseus*. *Proc. Natl Acad. Sci. USA* **97**, 5313–5316 (2000).
93. Wang, R.-L., Stec, A., Hey, J., Lukens, L. & Doebley, J. The limits of selection during maize domestication. *Nature* **398**, 236–239 (1999).
94. Felsenstein, J. Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.* **22**, 521–565 (1988).
95. Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).
96. 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).
97. Maddison, D. R. African origin of human mitochondrial DNA reexamined. *Syst. Zool.* **40**, 355–363 (1991).
98. Templeton, A. R. Human origins and analysis of mitochondrial DNA sequences. *Science* **255**, 737 (1992).
99. Hedges, S. B., Kumar, S., Tamura, K. & Stoneking, M. Human origins and analysis of mitochondrial DNA sequences. *Science* **255**, 737–739 (1992).
100. Ingman, M., Kaessmann, H., Pääbo, S. & Gyllensten, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713 (2000).
101. Mountain, J. L. Molecular evolution and modern human origins. *Evol. Anthropol.* **7**, 21–37 (1998).
102. Relethford, J. H. *Genetics and the Search for Modern Human Origins* (Wiley–Liss, New York, 2001).

Acknowledgements
We thank H. Innan and J. Pritchard for comments, and M. Tanaka, C. Wiuf and an anonymous reviewer for careful reading of the manuscript.

Online links

DATABASES
The following terms in this article are linked online to:
LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink
OdsH | Sod
MaizeDB: http://www.agron.missouri.edu
tb1

FURTHER INFORMATION
BATWING: http://www.maths.abdn.ac.uk/~ijw/downloads/downloads.htm
DnaSP: http://www.bio.ub.es/~julio/DnaSP.html
GENETREE: http://www.stats.ox.ac.uk/mathgen/software.html
LAMARC: http://evolution.genetics.washington.edu/lamarc.html
SITES: http://www.lifesci.rutgers.edu/~heylab
Access to this interactive links box is free online.

Genealogical trees, coalescent theory, and the analysis of genetic polymorphisms

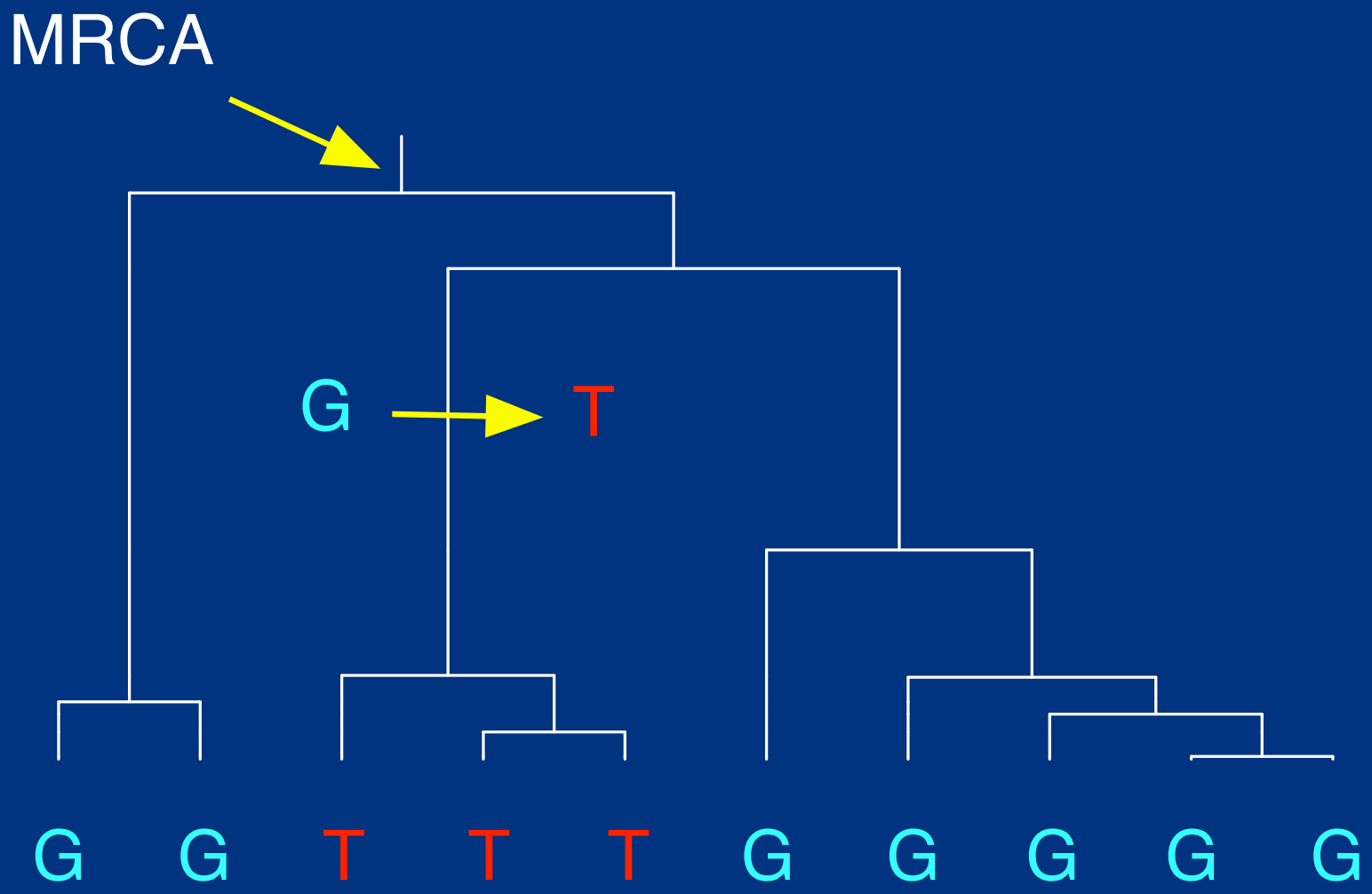
Magnus Nordborg

University of Southern California

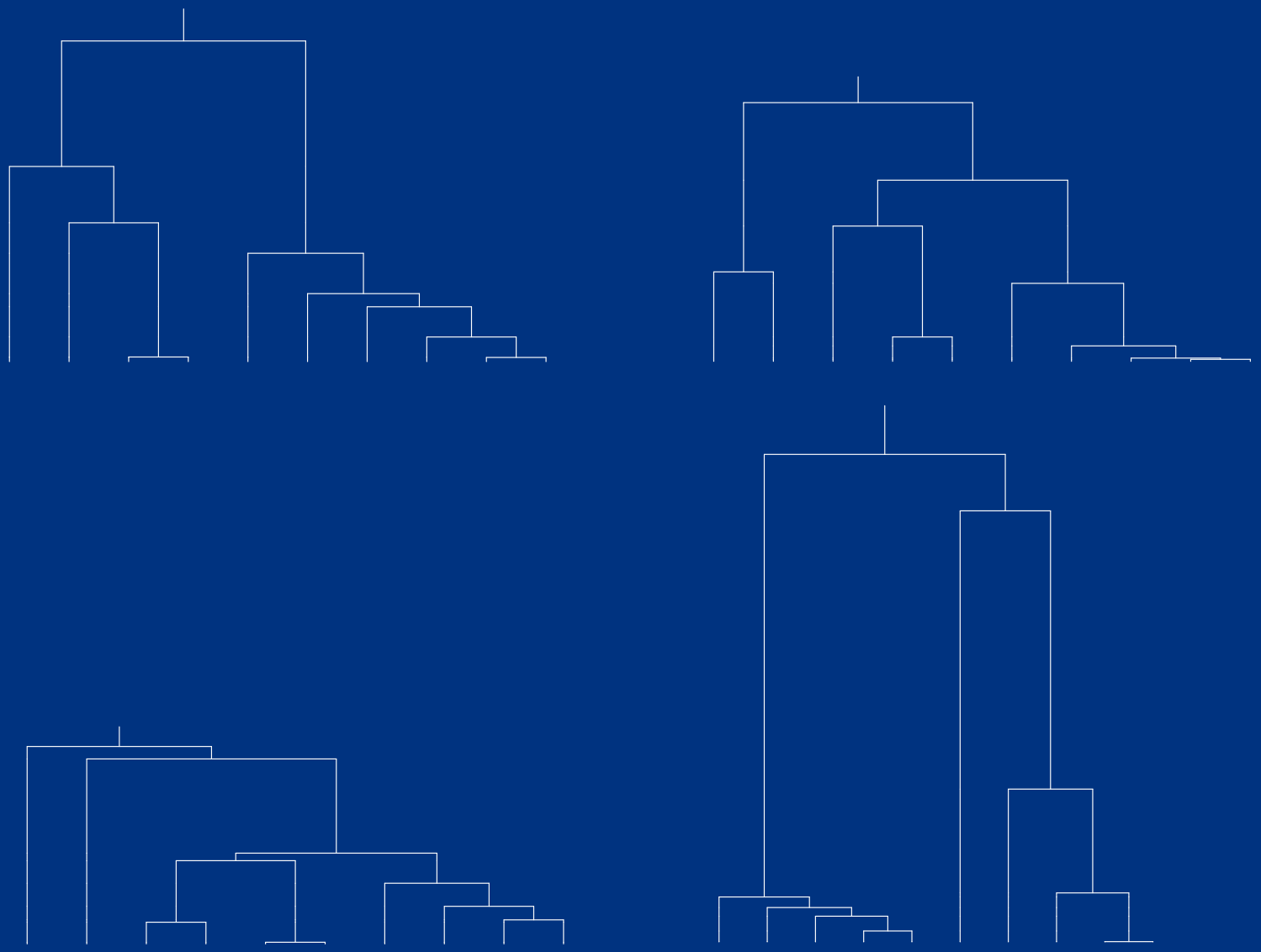
The importance of history

- Genetic polymorphism data represent the outcome of a single, highly complex, non-repeatable evolutionary history
- Traditional analysis methods cannot take this into account
- The stochastic process known as "the coalescent" presents a coherent statistical framework for analyzing genetic polymorphism data

The importance of history: mutations are random



The importance of history: trees are random

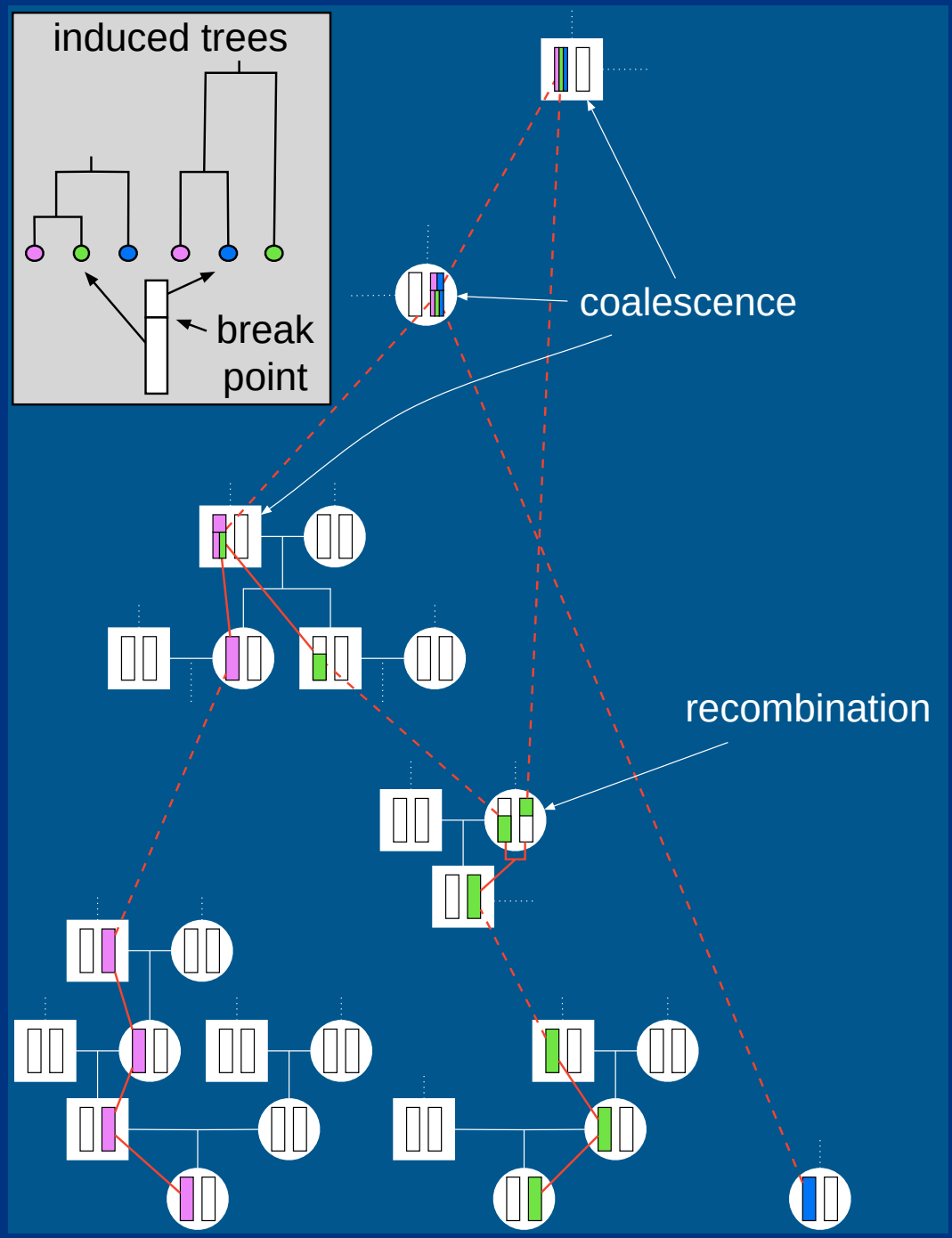


Modeling genetic polymorphism

At a minimum, models must include:

- coalescence (who begat whom, and when)
- mutation
- recombination

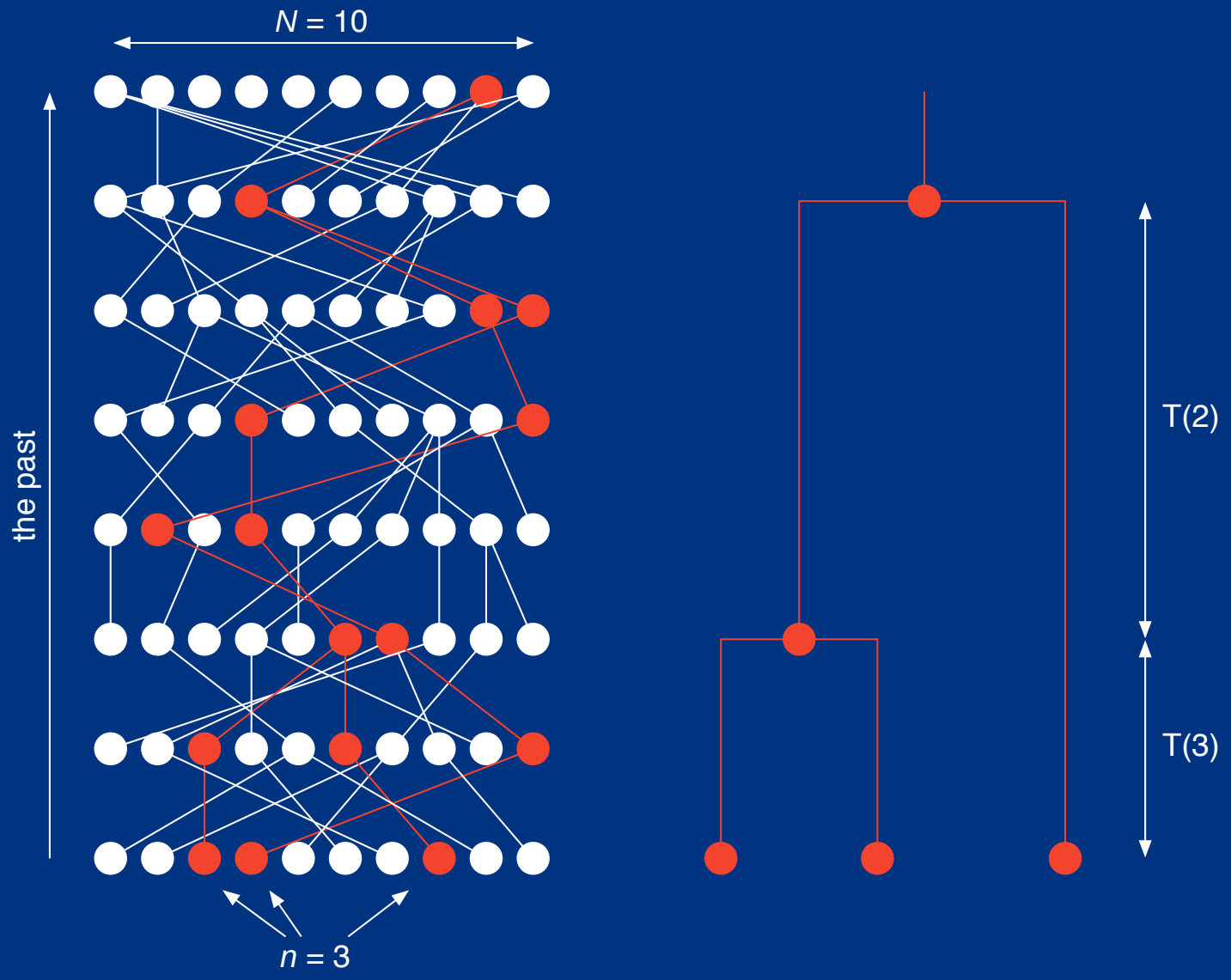
Recombination makes it possible for linked sites to have different genealogies



What is the coalescent?

- The coalescent is a stochastic process that is well-suited for modeling polymorphism data
- It is a natural extension to classical population genetics models

Coalescence: picking parents



The rate of coalescence

The rate at which lineages find each other depends on:

- The population size: the per-generation probability of coalescence is $\propto 1/N$
- The number of lineages: the rate of coalescence when there are k lineages is $\binom{k}{2}$
- A number of other demographic factors, such as inbreeding, age structure, and the variance in reproductive success

Because the per-generation probability of coalescence is on the order of $1/N$, we use a continuous-time approximation where time is measured in units of N generations

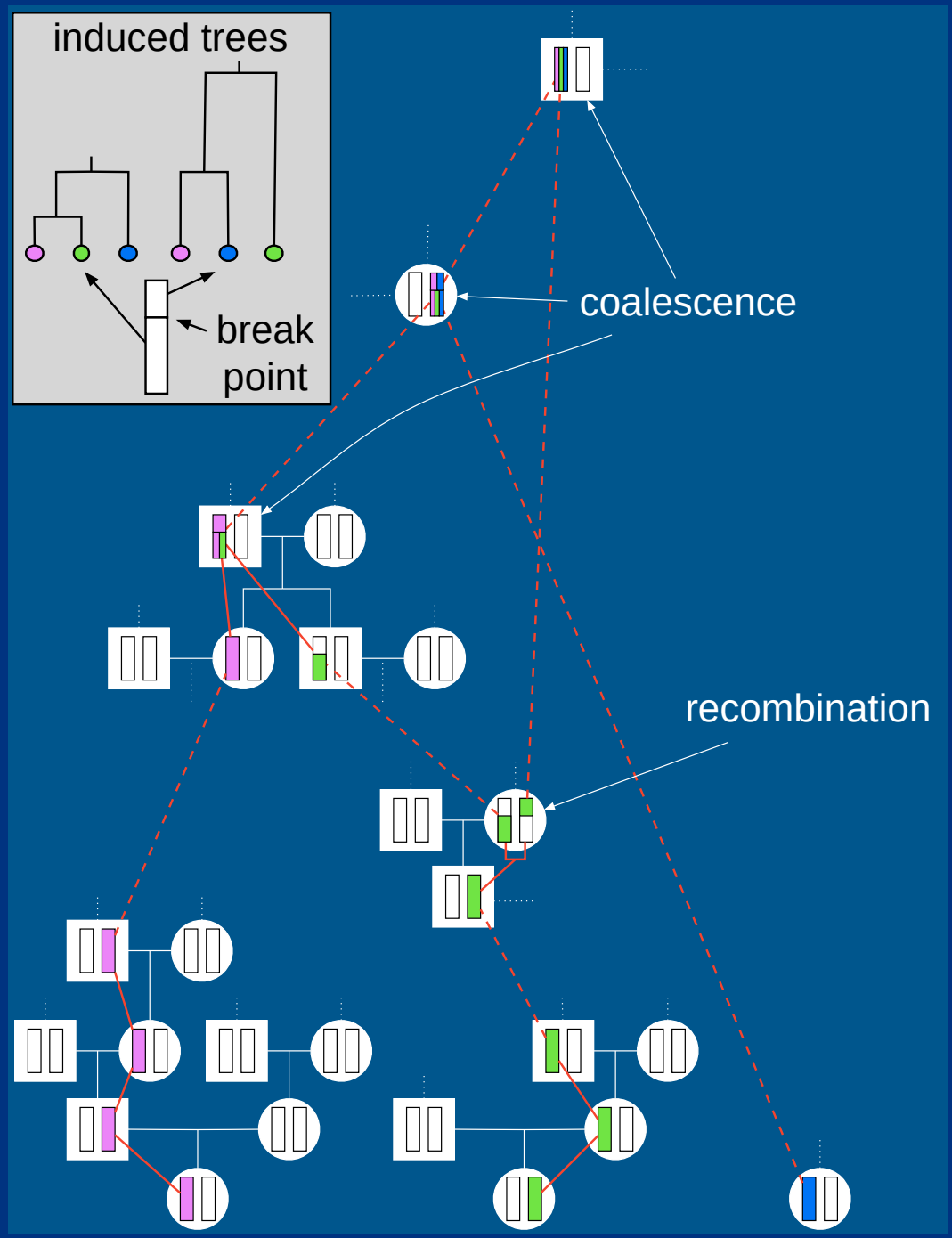
Mutation

- Selectively neutral mutations are added to the branches of the tree afterwards according to a rate that depends on the per-generation probability of mutation
- The expected number of mutations on a branch depends on its length — the expected number of mutations on the tree depends on the total branch length of the tree
- Any mutation model can be used

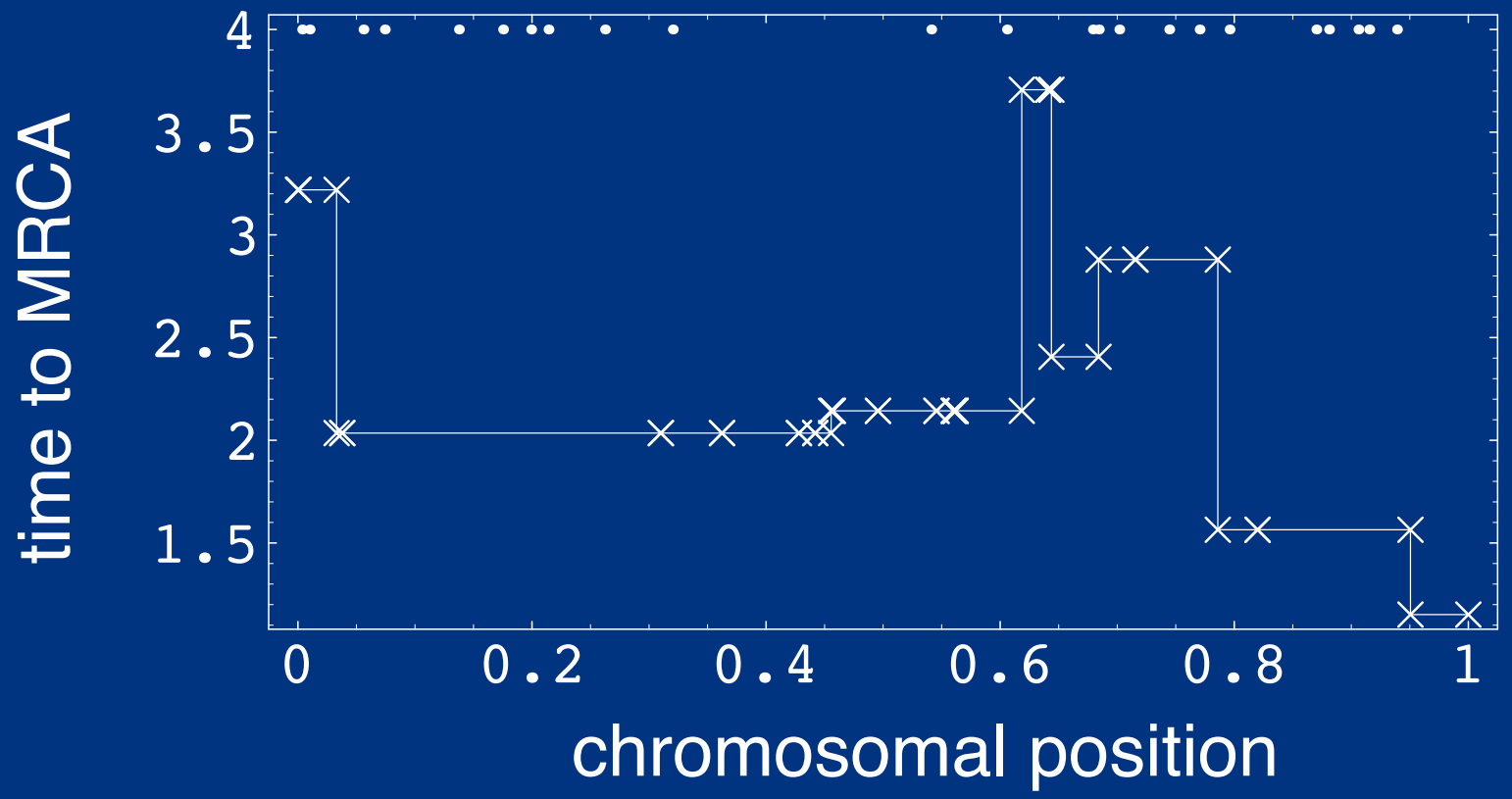
Recombination

- Recombination breaks up lineages according to a rate that depends on the per-generation probability of recombination
- There will be more recombination in the genealogy of a longer chromosomal segment
- Any recombination model can be used
- The coalescent with recombination generates a random graph — or a forest of trees

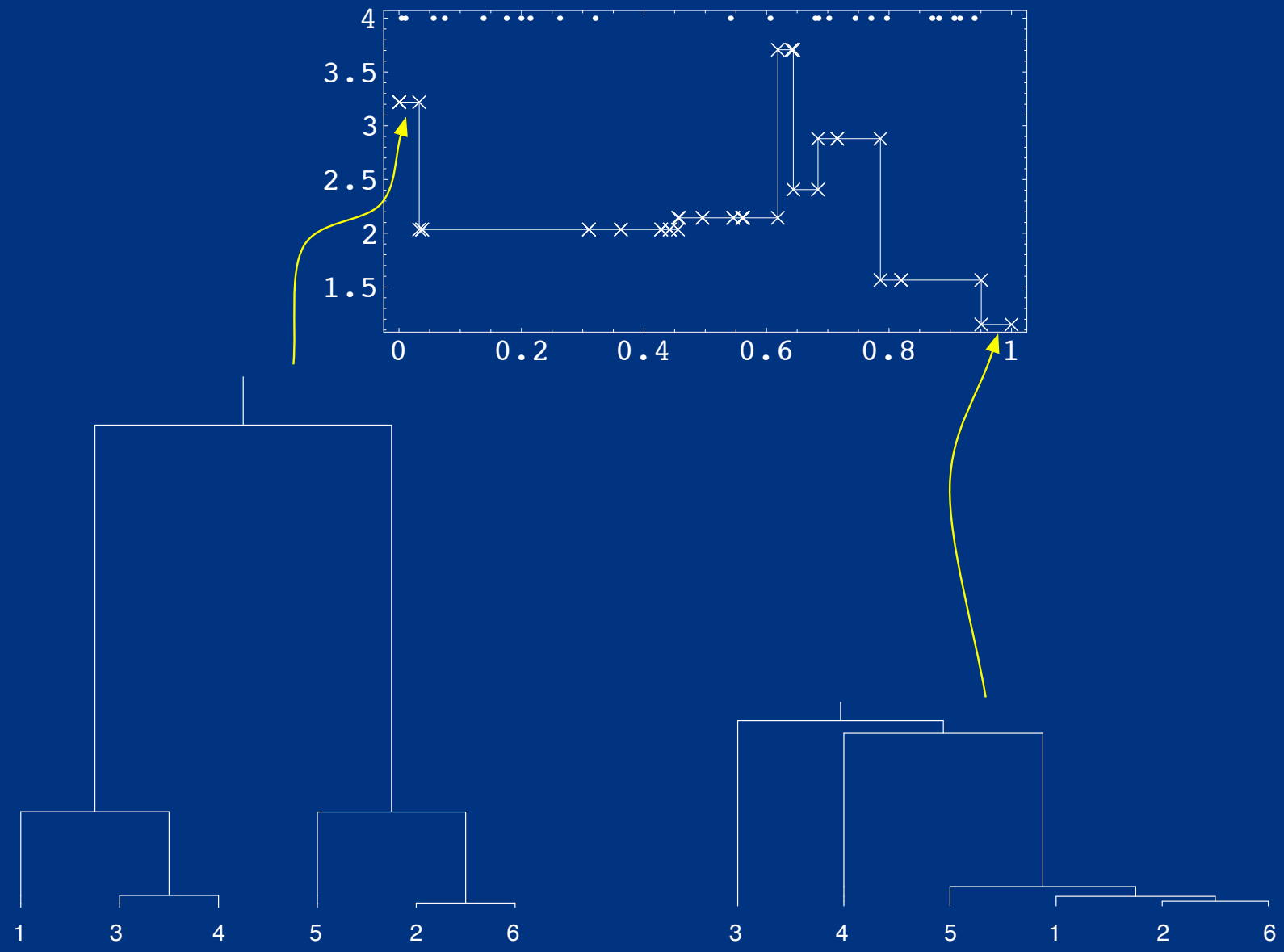
A graph or a forest. . .



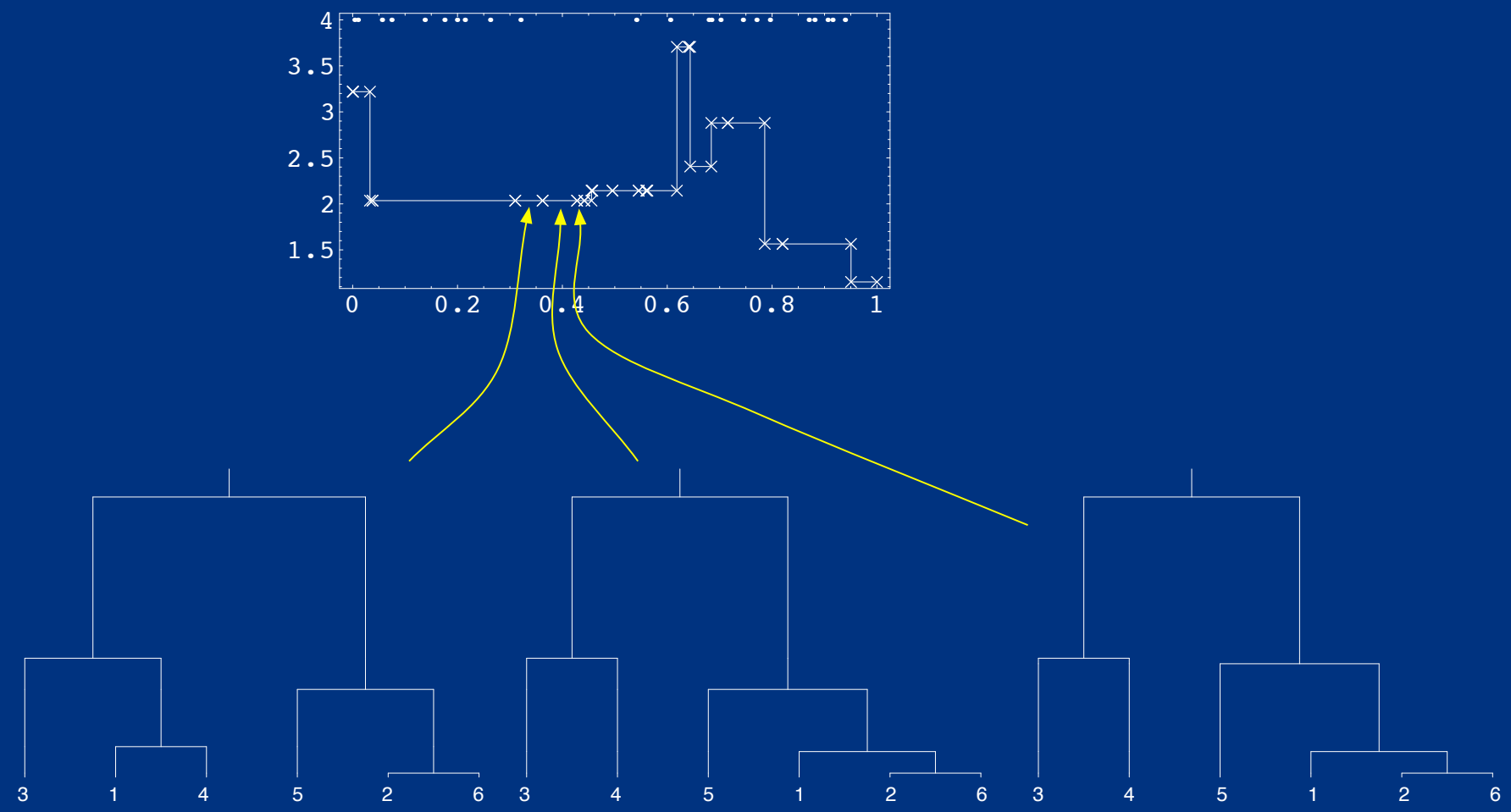
A walk through tree space



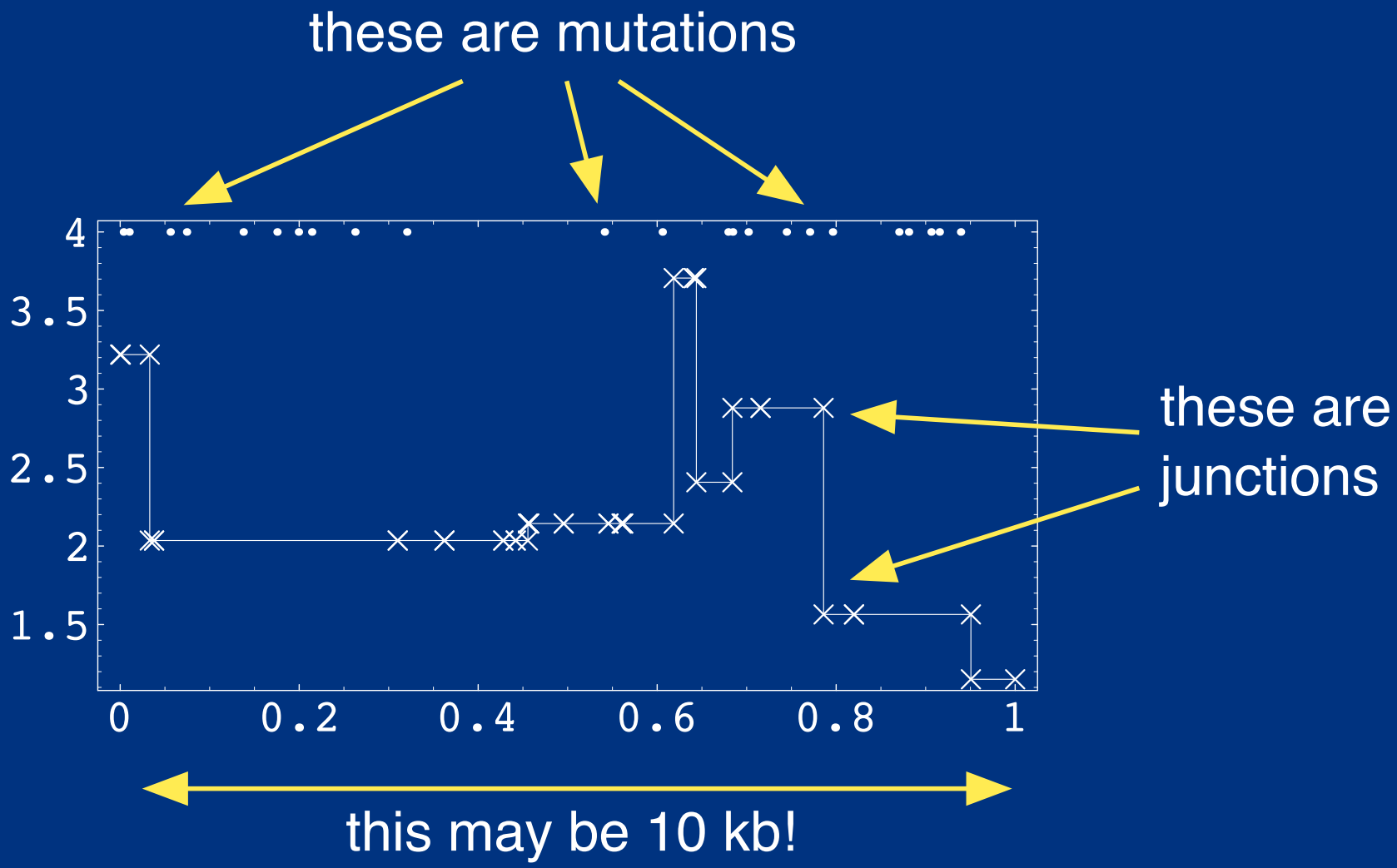
The trees are correlated



The trees are correlated



Recombination is common



Recombination is as common as mutation

- If 1 cM $\sim$ 1 Mb, then the probability of recombination per bp per generation is $\sim 10^{-8}$
- The probability of mutation per bp per generation is estimated to be *at most* 10^{-8}
- It follows that a sample of sequences will contain as many junctions as polymorphisms

Genealogical graphs can in general not be reconstructed

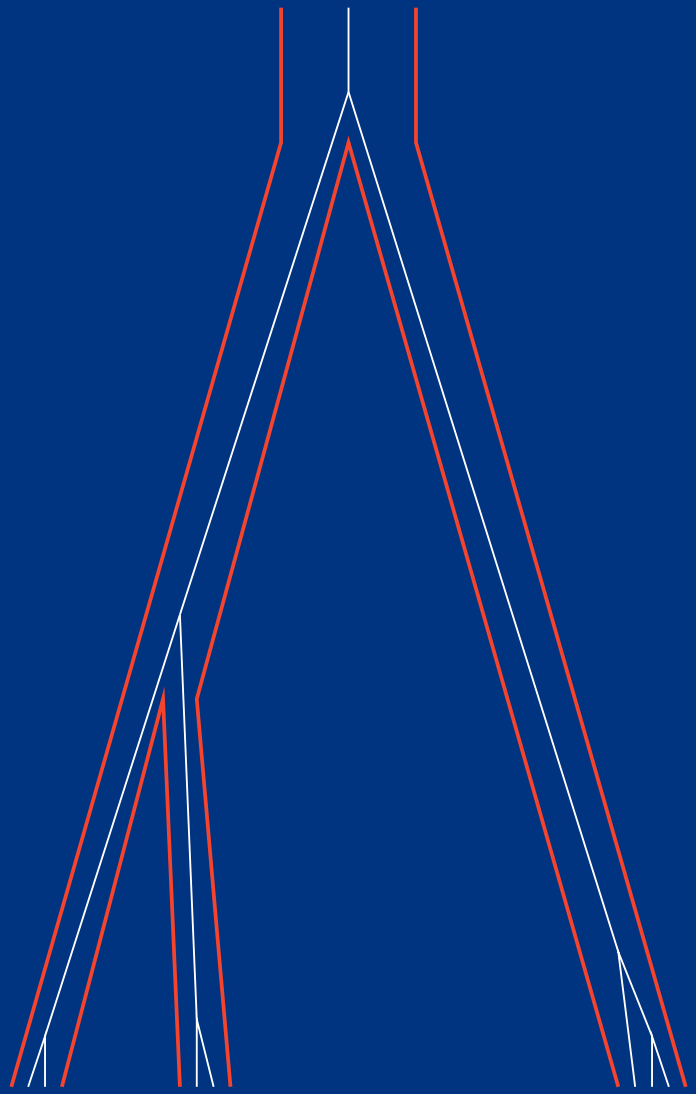
- Even with infinitely many polymorphisms, a substantial fraction of all junctions would not be detected
- In reality, there are clearly too few polymorphisms per junction to estimate the graph
- Remember: a phylogenetic algorithm will *always* reconstruct a tree, regardless of whether there exists a tree to be reconstructed...

We do not in general wish to reconstruct genealogical graphs

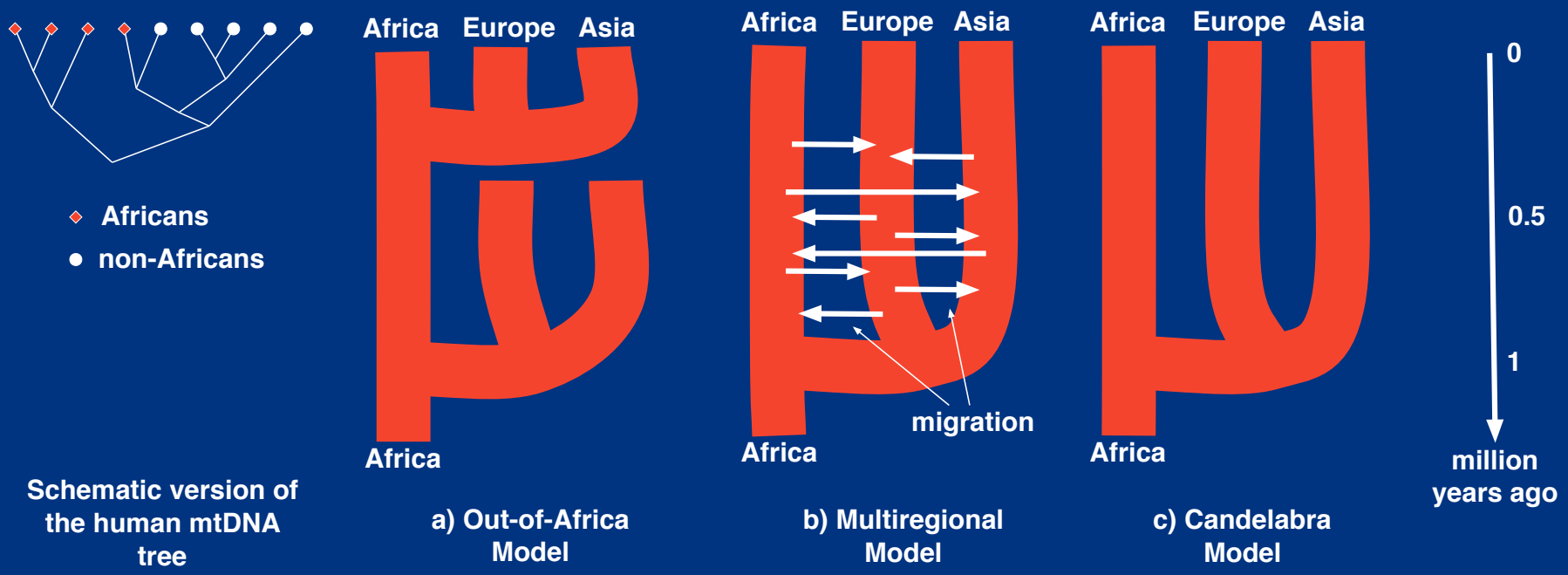
- Population genetics is not phylogenetics!
- Gene genealogies are of no interest *per se* — they are random outcomes of an underlying evolutionary process, and are of interest only insofar as they contain information about this process

Gene trees and species trees

Phylogenetic methods estimate species trees by estimating gene trees; they are appropriate if and only if the latter are strongly correlated with the former



Phylogenetic methods are not applicable to within-species data



- We must consider the likelihood of the data under alternative models

A likelihood framework

Phylogenetics:

$$L = \mathbb{P}(D|G, \mu)$$

Population genetics:

$$L = \sum_{G} \mathbb{P}(D|G, \mu)\mathbb{P}(G, \alpha)$$

Here D is the data, G the genealogy, μ the mutation model, and α the demographic model

Note that G is a *nuisance parameter* in population genetics

Uses of the coalescent

- A mathematical modeling tool
- A simulation tool for hypothesis testing and exploratory data analysis
- The basis for full likelihood inference

The simplicity and elegance of the coalescent process makes it a powerful modeling tool

At least for the standard coalescent, it is often possible to derive results analytically

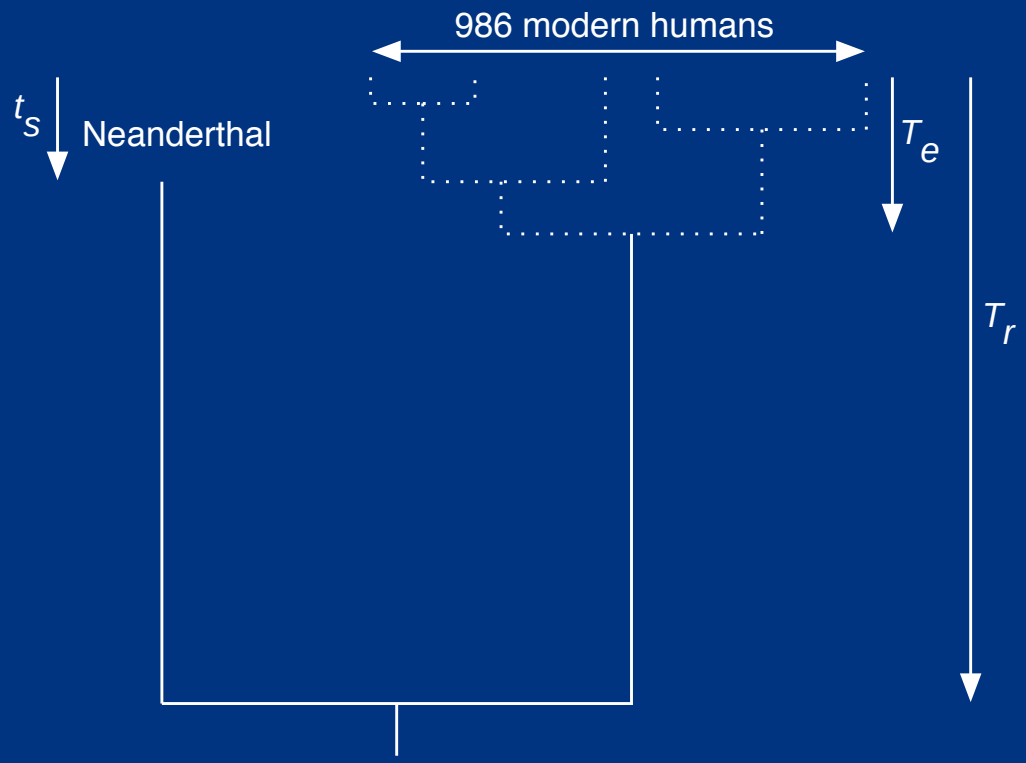
- Estimators and test, *e.g.*, Tajima's D statistic
- Illuminating theoretical results, *e.g.*, the probability that a sample of size n contains the MRCA of the entire population is

$$\frac{n-1}{n+1}$$

Almost any scenario can be simulated using the coalescent

- Coalescent simulations are enormously more efficient than classical methods
- Simulated data can be compared with real data — or used to evaluate the feasibility of a study before it is carried out

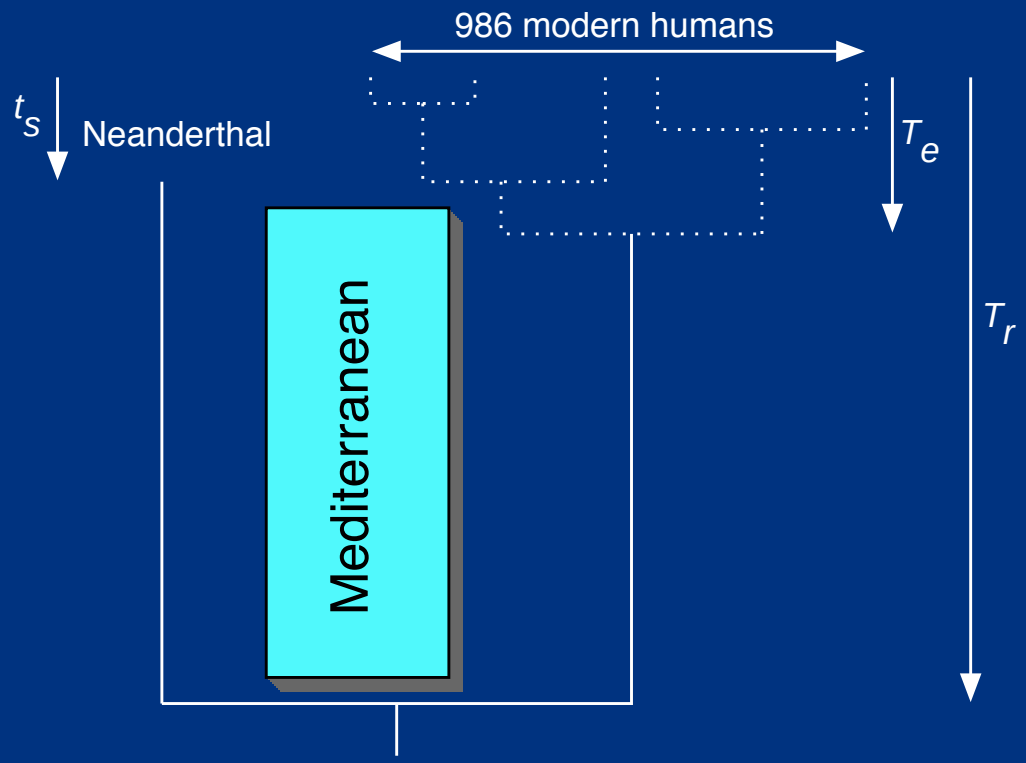
Example: ancient Neanderthal mtDNA



- Modern humans monophyletic
- $T_r > 4T_e$

Does this prove that Neanderthals and modern humans did not interbreed?

Example: ancient Neanderthal mtDNA



Assuming that they did interbreed, what is the probability of getting a tree like the one observed just by chance?

Coalescent simulations showed that this probability is high even for large amounts of interbreeding

Full likelihood analysis

- In principle possible
- In practice difficult
- Unless major breakthroughs are made, not likely to be applicable to genomic polymorphism data

What is the main insight from coalescent theory?

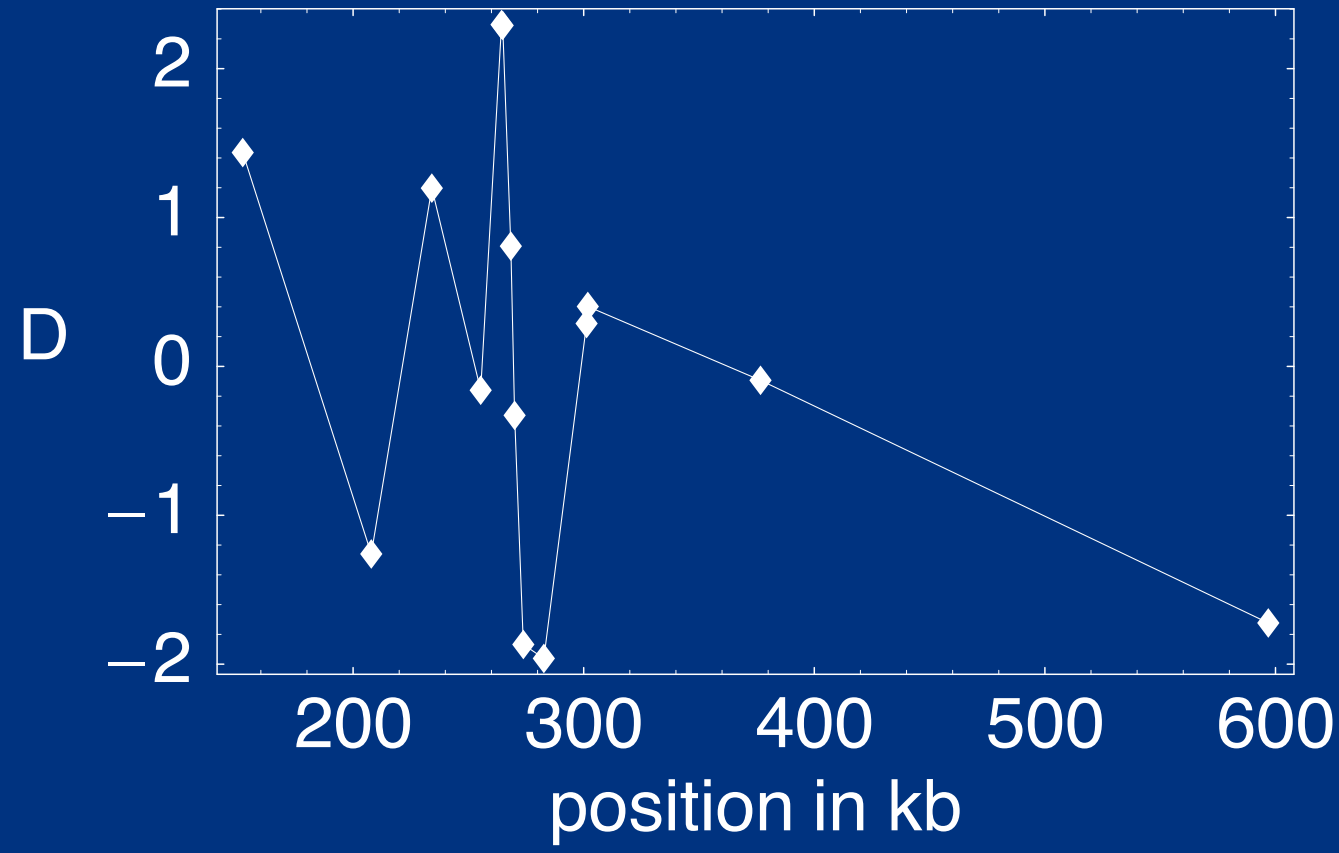
That very large numbers of loci are required to answer most questions!

Population Genomics is upon us!

- Data sets containing 100's and 1000's of loci already exist
- Within 10 years, it seems likely that whole-genome comparisons between species will be common, and that we will have whole genome sequences from 1000's of humans

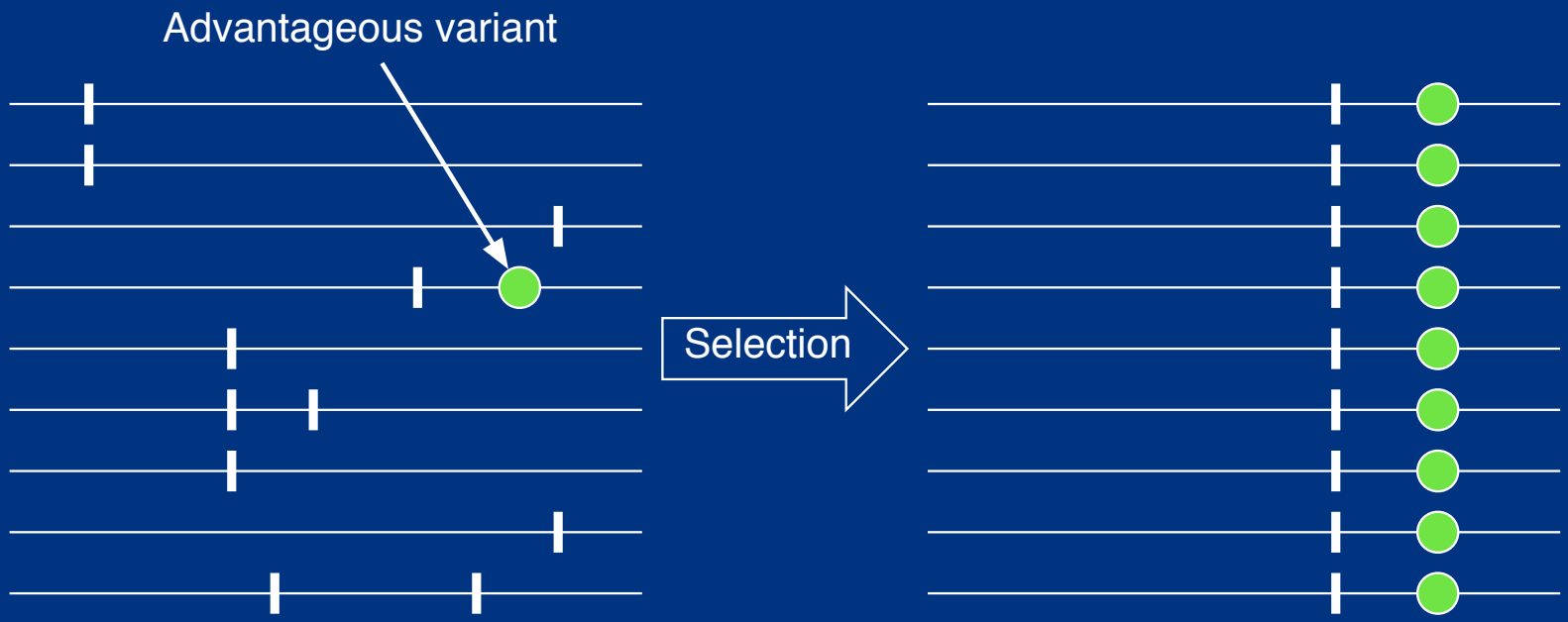
Less assumptions — more data

We will be able to use empirically estimated distributions of test statistics rather than theoretically predicted ones



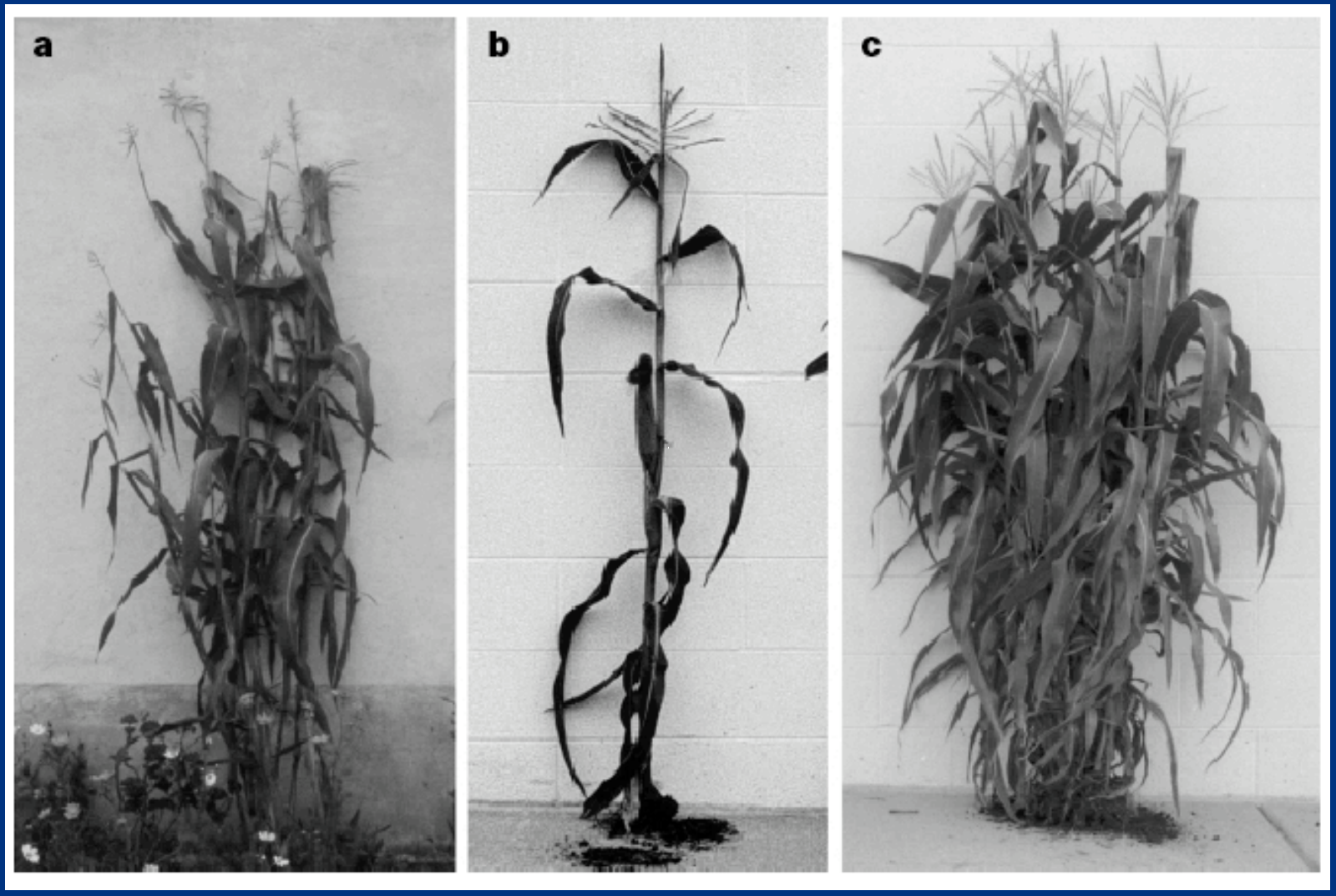
Selective sweeps

- Fixation of new alleles leaves a footprint in the pattern of genomic variation
- Can we find the genes that "make us human"



How many genes?

Teosinte to corn: $<$ 10,000 years; five genes?



teosinte maize maize with *tb1* mutation

What's the use polymorphism data?

- Whole-genome properties
  - demographic (*sensu lato*) history
  - molecular evolution
  - genetic mechanisms
- The history of individual loci — selection
  - divergence between human and other primates
  - traces of selection within the last million years

The history and future of multi-locus methods