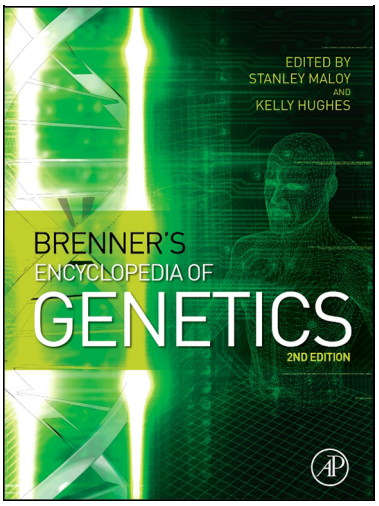




Marjoram P and Tavaré S, Coalescent. In: Stanley Maloy and Kelly Hughes, editors. Brenner's Encyclopedia of Genetics 2nd edition, Vol 2.
San Diego: Academic Press; 2013. p. 54–57.

# Coalescent

**P Marjoram,** University of Southern California, Los Angeles, CA, USA
**S Tavaré,** University of Southern California, Los Angeles, CA, USA; Cancer Research UK, Cambridge Research Institute, Cambridge, UK



This article is a revision of the previous edition article by C Neuhauser and S Tavaré, volume 1, pp 392–397, © 2001, Elsevier Inc.

## Glossary

**Association study** A study in which genetic polymorphism data are collected from a sample of individuals in an attempt to relate genetic variation to phenotypic variation (e.g., disease status).
**Bayesian statistics** A statistical analysis framework that allows for prior beliefs regarding a given inference problem, and which includes those prior beliefs within the inference process.
**Haplotyping** The process of determining haplotype information from a set of genotype data.
**Importance sampling** Statistical inference procedure for estimating properties of a given statistical distribution by sampling from another related distribution.
**Imputation** A procedure by which likely states of missing data are inferred.
**Markov chain Monte Carlo** A computationally intensive technique for sampling from probability distributions.
**The coalescent** A population genetics model that describes the ancestral history of a sample of individuals drawn from an evolving population.
**Urn model** A model in which a problem is described by drawing balls of different colors from an imaginary urn. Depending upon context, the balls might represent chromosomes, mitochondrial DNA, and so on.
**Wright–Fisher model** A mathematical model for the evolution of a population of randomly mating individuals.

## Introduction

The existence of a rapidly growing number of large-scale, population-based studies of molecular variability, often obtained as random samples of DNA sequences or as samples of single-nucleotide polymorphisms, offers great potential for improving our understanding of the relationships between genetic variability and phenotype. Because the individuals in the sample are related, these data are highly dependent; understanding the nature of this dependence is crucial for the analysis of the variability in the sample.

The coalescent, introduced by Kingman in 1982, borrowed intuition from pedigree-based analyses, observing that while dependence relationships between members of randomly drawn samples are typically not known, they do still exist. It is a probabilistic model of those unobserved relationships and of the occurrence of mutations to the ancestry of the sample. The use of genealogical or coalescent methods is now central to efficient analysis of genetic data, providing a natural framework for estimation of, and inference about, evolutionary parameters.

## The Ancestral Process – The Neutral Case

The genealogy of a sample drawn from a population can be visualized as a coalescing tree. A realization is shown in Figure 1. A tree that corresponds to a sample of size $n$ has $n$ tips and one root. The root corresponds to the most recent common ancestor.

The coalescent provides a formal mathematical description of this genealogy and of the occurrence of mutations on it. In its simplest, neutral form, it is derived by assuming that the population is haploid and of fixed size $N$ individuals. Furthermore, we assume that the population evolves forward in time according to the discrete-time Wright–Fisher model. In this model, each individual has a binomial number of offspring, conditioned on keeping the total population size constant.

When the population size $N$ is large compared with the sample size, the genealogy of a sample of size $n$ can be approximated by a continuous-time Markov chain $A(t)$ in which time $t$ is measured in units of $N$ generations. The process starts from $A(0) = n$ and goes through the states $n, n-1, \ldots, 2, 1$. A value of $A(t) = j$ means that the sample had $j$ distinct ancestors time $t$ ago. The amount of time $T_j$ for which there are $j$ ancestors is exponentially distributed with mean $2/(j(j-1))$, and these times are independent of one another. This Markov chain $A(t)$ is often called the coalescent process.

A characteristic of the neutral genealogy for fixed population size is that the last two branches dominate the height of the tree. This can be seen by comparing the expected coalescence time of two branches, $\mathrm{ET}_2$, and the expected time to the most recent common ancestor, $T_{\mathrm{MRCA}}$. The expected time until two ancestors coalesce is 1, which is more than half of the total expected time to the most recent common ancestor, regardless of the sample size.

In the neutral case, demography and the mutation process can be separated. Thus, to obtain a sample of size $n$, one can first construct its genealogy and then superimpose the mutation process on that genealogy. This provides an extremely efficient way to simulate observations from complicated demographic and mutation scenarios.

We assume the simplest mutation process in which mutations occur independently to all genes with probability $u$ per gene per generation. If time is scaled in units of $N$ generations and if

$$\lim_{N\to\infty} 2Nu = \theta$$

 doi:10.1016/B978-0-12-374984-0.00274-6

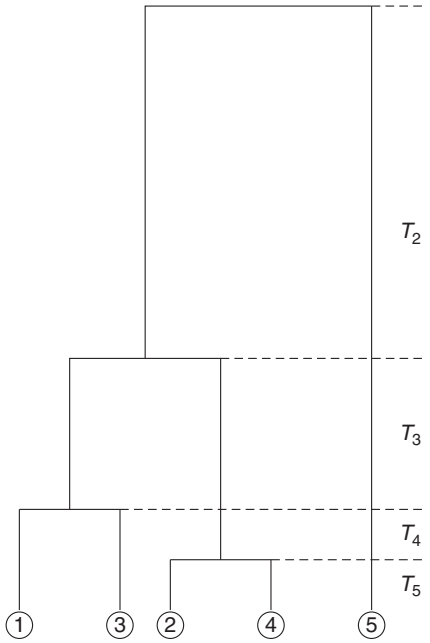


**Figure 1** Coalescent tree of sample of five individuals.

then mutations occur along the branches of the coalescent process according to a Poisson process with rate $\theta/2$, independently in each branch of the coalescent.

The distribution of the total number of mutations in the sample since their most recent common ancestor follows readily. Given the total length $L$ of the branches in the tree, which is

$$L = \sum_{j=2}^{n} jT_j$$

the total number of mutations in the tree follows a Poisson distribution with mean $\theta L/2$.

## Robustness

The coalescent is remarkably robust. It provides a good approximation for a large class of reproduction models when the population size $N$ is large relative to the sample size $n$. This class includes both discrete-time models in which generations do not overlap and continuous-time models in which generations overlap.

Furthermore, genealogies can be formulated for diploid populations. In the neutral case when mating is random (i.e., a panmictic population), diploidy simply means that the number of genes is doubled: if the population size is $N$, then the number of genes is $2N$. The genealogy in the diploid case is then the same as in the haploid case with $N$ replaced by $2N$.

## Varying Population Size

It is straightforward to incorporate deterministically varying population size into the ancestral process. This only affects the coalescence rate and is therefore the same for both the neutral and the selective case.

The effect of a growing population can be quite dramatic. For instance, if the population has grown exponentially, the resulting graph is stretched near the present time and compressed in the past (i.e., near the root). The resulting graph resembles a star phylogeny in the neutral case.

## Recombination

To describe the genealogy of two linked regions, or loci, we assume that recombination occurs independently in each offspring. In each generation, with probability $1-r$, each offspring independently inherits both regions from the same chromosome; with probability $r$, the genes are inherited from different chromosomes (i.e., a recombination event occurred).

The resulting Markov chain, known as the ancestral recombination graph (ARG), can be described graphically, which shows the lineages of each individual in the sample. Following a lineage backward in time on this graph, recombination events occur at rate $\rho/2$, where $2Nr=\rho$. At such times, the lineage splits and results in a branching event. By convention, branches that correspond to the left locus are drawn to the left and branches that correspond to the right locus are drawn to the right at branching points. Common ancestry is again represented by the coalescing of branches. An example is given in Figure 2.

The ancestry of each locus can be traced separately by following the paths to the left for the $L_1$ locus and to the right for the $L_2$ locus at each branching point. It follows that the ancestry of each locus is given by the neutral coalescent process, but these marginal coalescent trees are of course not independent of one another. The ancestral graph can be generalized in a natural way to describe multiple loci.

## Selection

When natural selection is incorporated into the model, the ability to separate the demography and the mutation process

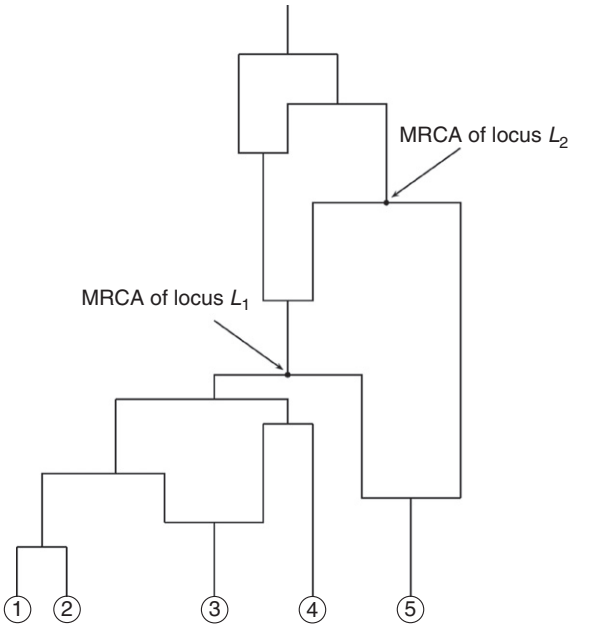


**Figure 2** Two-locus ancestral recombination graph for sample of five individuals. MRCA, most recent common ancestor.

is lost. This is because reproductive success depends on the allele of an individual, and it results in a more complicated structure, the ancestral selection graph, which has similar topological structure to the ARG.

However, if selection is sufficiently strong, one can again separate demography and mutation, at least approximately. The embedded genealogy then becomes approximately a simple time change of Kingman's neutral coalescent. The reason for this is that under strong selection, the population dynamics are on a much faster timescale than coalescing events.

In cases where this separation of timescale occurs, the ancestral process can be modeled as a change in the effective population size. In particular, this says that not only are the expected times between coalescing events a time change relative to the neutral case but the distribution of the coalescing events is the same as in the neutral case except for the timescale.

## Population Subdivision

The assumption of panmixia can be replaced by the assumption that the population is geographically structured. The simplest case is that of a subdivided population in which the population consists of a finite number of islands, each populated by a subpopulation. Reproduction on each island follows the Wright–Fisher model (possibly with selection). Each generation, a proportion $m_{ij}$ of the offspring on island $i$ migrates to island $j$, regardless of their genotype. Various more general versions of this scenario are also possible.

The effect of population subdivision compared with the panmictic case is a compression of the coalescent near the tips of the tree due to the smaller sizes of the subpopulations. However, further back in the past, provided the migration rate is small enough, the branches are extended since lineages have to be on the same island in order to coalesce. The coalescence time in the subdivided population shows much greater variance than that in the panmictic case.

## Approximations

The size of modern data sets oftentimes requires approximation of the underlying models in order to ensure computational tractability. The coalescent is no exception to this. Li and Stephens introduced a delightful approximation based upon the so-called Urn models; other approaches have also emerged using more explicit approximations, such as the sequential Markovian coalescent, encoded with the Markovian Coalescent Simulator (MACS) algorithm (see Model-Based Analyses).

## Inference

An important use of the coalescent arises when estimating population parameters such as mutation and recombination rates. A number of approaches have been proposed for this purpose, including those based on the behavior of summary statistics (e.g., the number of segregating sites observed in a sample of DNA sequences is often used to estimate the mutation rate). Full likelihood methods and Bayesian approaches are currently of great interest, particularly as they provide an inferential framework for mapping disease genes by linkage disequilibrium mapping and by haplotype sharing. Importance sampling and Markov chain Monte Carlo approaches have been proved useful in this context.

An important question in this context is, 'can the coalescent be used to produce data that appear to mimic those found in human populations?' Schaffner *et al*. showed that the flexibility of the coalescent was sufficient to well-approximate data seen in human populations.

## Model-Based Analyses

The coalescent is now used as the basis of a number of computationally intensive methods for data analysis. We give examples here. References to websites from which these methods can be downloaded are listed below.

*Imputation and haplotyping*. Software such as PHASE, fastPHASE, and IMPUTE use approximations to the coalescent that address the important issue of imputing likely values for missing genotype data or determining the phase of genotype data.

*Estimation of evolutionary parameters*. LAMARC and GENETREE use coalescent trees to estimate mutation, migration, and recombination rates. Approximate Bayesian computation (cf. Beaumont) is sometimes a useful approach.

*Association studies*. An area of great interest is the analysis of genome-wide association studies, designed to detect genetic loci that are associated with disease status. Several such tools use the coalescent, either explicitly or implicitly. For example, MARGARITA, COLDMAP, and LAMARC.

*Data simulation*. Many studies demand the simulation of genetic data. The most popular simulation tool for genotype data is Hudson's ms program. For larger regions, programs such as GENOME or MACS can be used.

*See also:* Evolutionary Trees; Fisher, R.A.; Gene Trees; Genetic Drift; Genetic Variation; Haplotype; Linkage Disequilibrium; Neutral Theory; Population Genetics; Population Substructure; Wright, Sewall.

## Further Reading

Beaumont MA (2010) Approximate Bayesian computation in evolution and ecology. *Annual Reviews of Ecology, Evolution, and Systematics* 41: 379–405.

Hein J, Schierup MH, and Wiuf C (2005) *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford: Oxford University Press.

Kingman JFC (1982) On the genealogy of large populations. *Journal of Applied Probability* 19A: 27–43.

Li N and Stephens M (2003) Modelling linkage disequilibrium, and identifying recombination hotspots using SNP data. *Genetics* 165: 2213–2233.

Marjoram P and Tavaré S (2006) Modern computational approaches for analysing molecular genetic variation data. *Nature Reviews Genetics* 7: 759–770.

Nordborg M (2008) Coalescent theory. In: Balding DJ, Bishop MJ, and Cannings C (eds.) *Handbook of Statistical Genetics*, 3rd edn. New York: John Wiley & Sons, Inc. doi:10.1002/9780470061619.ch25.

Schaffner SF, Foo C, Gabriel S, *et al*. (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Research* 15: 1576–1583.

Tavaré S (2004) *Lectures on Probability Theory and Statistics: Ecole d'Eté de Probabilités de Saint-Flour XXXI – 2001 (Lecture Notes in Mathematics, vol. 1837, pp. 1–188)*. Berlin: Springer.

Wakeley J (2008) *Coalescent Theory: An Introduction*. Greenwood Village, CO: Roberts & Company.

## Relevant Websites

http://www.stanford.edu/group/hadlylab/ssc/index.html – BayeSSC.
http://stephenslab.uchicago.edu/software.html#fastphase – fastPHASE.
http://www.stats.ox.ac.uk/~griff/software.html – GENETREE.
http://www.sph.umich.edu/csg/liang/genome/ – GENOME.
http://mathgen.stats.ox.ac.uk/impute/impute.html – IMPUTE.
http://evolution.genetics.washington.edu/lamarc/index.html – LAMARC.
http://www-hsc.usc.edu/~garykche/ – MACS.
http://www.sanger.ac.uk/resources/software/margarita/ – MARGARITA.
http://home.uchicago.edu/rhudson1/source/mksamples.html – ms.
http://cmpg.unibe.ch/software/simcoal2/ – SimCoal2.