Coalescent Theory

Intermediate article

Simon Tavaré, *University of Southern California, Los Angeles, California, USA*

Coalescent theory is a very useful tool in the interpretation of genomic variation data.

Article contents

Introduction

One of the most important challenges facing modern biology is how to make sense of genetic variation. Understanding how genotypic variation translates into phenotypic variation, and how it is structured in populations, is fundamental to our understanding of evolution. Understanding the genetic basis of variation in phenotypes such as disease susceptibility, drug response and aging are of primary importance to human geneticists. Until very recently, our knowledge about human genetic variation was limited to surveys of small numbers of markers or to sequence surveys of individual genes, but this is rapidly changing because of technological advances. The most dramatic examples to date are provided by Perlegen Sciences Inc., who resequenced 20 copies of chromosome 21 (Patil *et al*., 2001), and by Genaissance Pharmaceuticals Inc., who studied haplotype variation and linkage disequilibrium across 313 human genes (Stephens *et al*., 2001). Although these studies represent staggering amounts of data, they are only the tip of the genomic iceberg; several other genomic-scale studies are already under way. Interpretation of genome-wide polymorphism data has become a pressing priority. In this article, I describe the basics of *coalescent theory*, a useful quantitative tool in this endeavor.

What is a Coalescent?

Much of the basic quantitative theory of population genetics, as developed by Wright, Fisher, Haldane, Kimura and others, has revolved around stochastic models that describe the evolution, forward in time, of a collection of gene frequencies. Provine (2001) provides a historical perspective. A basic tool in this approach is *diffusion theory*, which arises in the context of models for gene frequency evolution in large populations. This theory is described in Ewens (1979) and reviewed further in Tavaré (1984) and Neuhauser (2001). In the late 1970s, it became clear that this *prospective* approach was not ideal for interpreting molecular variability obtained from *samples* of chromosomal regions. Explaining the structure of extant variation is more naturally posed as a problem about the *retrospective* behavior of the population that was sampled (cf. Ewens, 1990). In particular, the ancestral relationships among the chromosomal regions in the sample are important, and this leads directly to the study of ancestral processes and the coalescent.

Imagine first tracing back the ancestry of a sample of nonrecombining segments of deoxyribonucleic acid (DNA) – for definiteness, think of the mitochondrial D loop. Each segment in the sample is a copy of that segment in its parent, each of those parental regions is a copy of the segment in its parent, and so on back in time. Sometimes these ancestors will be shared by members of the sample; when two or more sample segments first share a common ancestor, we say a coalescence event has occurred. Proceeding back in time, we notice that the number of distinct ancestral segments decreases, until eventually all the members of the sample are traced back to their most recent common ancestor (MRCA). Another way to think of this process of coalescences is as a random rooted tree: the leaves of the tree represent the sample segments, branches coalesce whenever the segments on those branches share a common ancestor, and the root corresponds to the MRCA. The tree is random because a different sample will have a different tree. An illustration is given in **Figure 1**. An ancestral tree such as that in **Figure 1** is described by its topology – the shape of the tree (equivalently, a listing of who is related to whom) – and by the distribution of the times at which coalescence events occur.

Kingman (1982a), Hudson (1983) and Tajima (1983) pioneered the development of coalescent theory in the context of large populations. Consider first a population of N segments, and suppose that the numbers $\nu_1, \ldots, \nu_N$ of 'offspring segments' born to segments $1, \ldots, N$ are exchangeable (so in particular they all have the same marginal distribution; Cannings (1974)). The constant size ensures that the mean number of offspring segments is $\mathbb{E}\nu_1 = 1$. We further assume that the generations behave independently and that the offspring distributions are identical from generation to generation. Kingman (1982b) showed

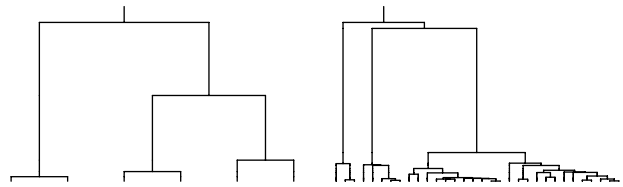

Figure 1 Coalescent trees for samples of size 6 and 32.

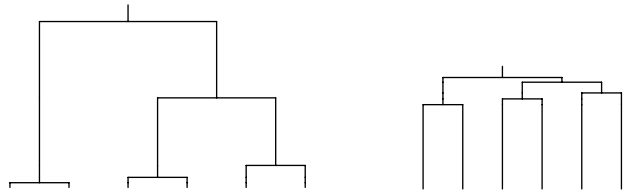

Figure 2 Coalescent tree of a sample of size 6 (constant population size in the left panel, exponentially growing population in the right panel).

that if the variance of v_1 has a finite positive variance σ^2 and time is rescaled in units of $\sigma^{-2}N$ generations, then in the limit as $N \to \infty$ the ancestral tree of a sample of n segments can be described in a very simple way:

(K1) The tree topology is uniform: first choose a pair of leaves and join them. This produces $n-1$ ancestors, from which a random pair is joined, and so on. Note that, on this timescale, only two individuals can find a common ancestor at the same time; the tree is bifurcating.

(K2) The times $T_n, T_{n-1}, \ldots, T_2$ during which the sample has $n, n-1, \ldots, 2$ distinct ancestors are independent and exponentially distributed random variables with means $\mathbb{E}T_j = 2/(j(j-1))$.

The condition that σ^2 has a positive finite limit excludes the case in which $\upsilon_i \equiv 1$ for all i (for which $\sigma^2 = 0$ and there is no interesting genealogy), and the case in which a random segment has all the offspring (for which $\sigma^2 = N-1$, and everyone is traced back to a common ancestor in a single generation). The classical Wright–Fisher model is the case in which the υ_i have a symmetric multinomial distribution (and so $\sigma^2 = 1$). There is now a substantial literature concerning approximations to the genealogy of exchangeable and related models (cf. Möhle, 1998), some of which are discrete-time processes (Möhle and Sagitov, 2001). In the remainder of this review, we focus on Kingman's continuous-time approximation. In particular, we assume unless otherwise stated that $\sigma^2 = 1$, noting only that the value of σ^2 is important in translating results from the coalescent timescale to one measured in generations or years.

From (K2) above, it follows that the height W_n of the coalescent tree of a sample of size n is:

$$\mathbb{E}W_n = \mathbb{E}(T_n + \cdots + T_2) = 2(1 - 1/n)$$

corresponding to about $2N$ generations. The variance of W_n follows from (K2) as well; in particular $1 < \mathrm{Var}\, W_n \leq \mathrm{Var}\, W_\infty \approx 1.16$. Coalescent trees tend to be very variable, most of this variability being attributable to the last coalescence in the tree; on average over half the height of the tree comes from T_2. This phenomenon is illustrated by the right-hand panel in **Figure 1**.

Relatives of the coalescent

Coalescents have now been developed to address many other models of demography, including variable population size, subdivision and migration. The effects of variable population size are easiest to describe. Consider a population of current size N having relative size $f(t)$ time t ago, where time is measured in units of N generations once more. If at time t the ancestral tree has k ancestors in it, then the rate at which coalescence events occur is $k(k-1)/2f(t)$. In an exponentially growing population, the effect is to stretch the tree near the leaves and compress it near the root. Intuitively, few coalescences happen while the population is large. If the population has grown from a small size in the past, the tree will look star-shaped (**Figure 2**). For a comprehensive review of other aspects of demography and segregation, see Nordborg (2001).

Effects of mutation

The effects of neutral mutations can be superimposed on the genealogical tree. In the original discrete-time model, suppose that mutations arise with probability u per segment per generation. The expected number of mutations arising along a lineage of length g generations is therefore gu, so that with time measured in units of $\sigma^{-2}N$ generations, we have

$$gu = \frac{\sigma^2 g}{N}\frac{Nu}{\sigma^2} \sim t\frac{\theta}{2}$$

where we have defined $\theta = 2Nu/\sigma^2$. It follows that as long as mutation is occurring at a rate proportional to the reciprocal of the population size, there is a balance between coalescence events and mutations. In this limiting regime, mutations are poured down the branches of the coalescent tree at rate $\theta/2$, independently for each branch of the tree.

The effects of these mutations can be modeled in many ways, depending on the data at hand. Of particular relevance to single-nucleotide polymorphism data is the so-called *infinitely-many-sites* model. Model the chromosomal segment of interest as a unit interval, and identify mutations with labels chosen independently, uniformly and at random from (0,1). All segments that share the ancestor in which a particular mutation occurred must have the mutant allele at that site, and the remaining segments have the ancestral type. Notice that, in this case, no location in the segment can have a mutation more than once. To model mutational hot spots across a chromosomal segment, one need only change the distribution of the labels to something nonuniform. Another case to consider is that of DNA sequence data. One needs to specify a mutation model to describe the effects of mutations in the s base pair region. One simple model chooses a site at random to mutate, and specifies a matrix of transition probabilities that give the probability that the current base is substituted by another. A popular choice is Felsenstein's model (cf. Thorne *et al*., 1992). Rate heterogeneity can be incorporated as before.

Ancestral Recombination Graph

Arguably, the most important aspect of coalescent theory is its treatment of the effects of recombination. This is particularly important in the light of genome-wide variation studies and applications to mapping genes. This theory, pioneered by Hudson (1983), Kaplan and Hudson (1985) and Griffiths (1991), is most easily described by considering a small pedigree, as shown in **Figure 3**, adapted from Nordborg and Tavaré (2002).

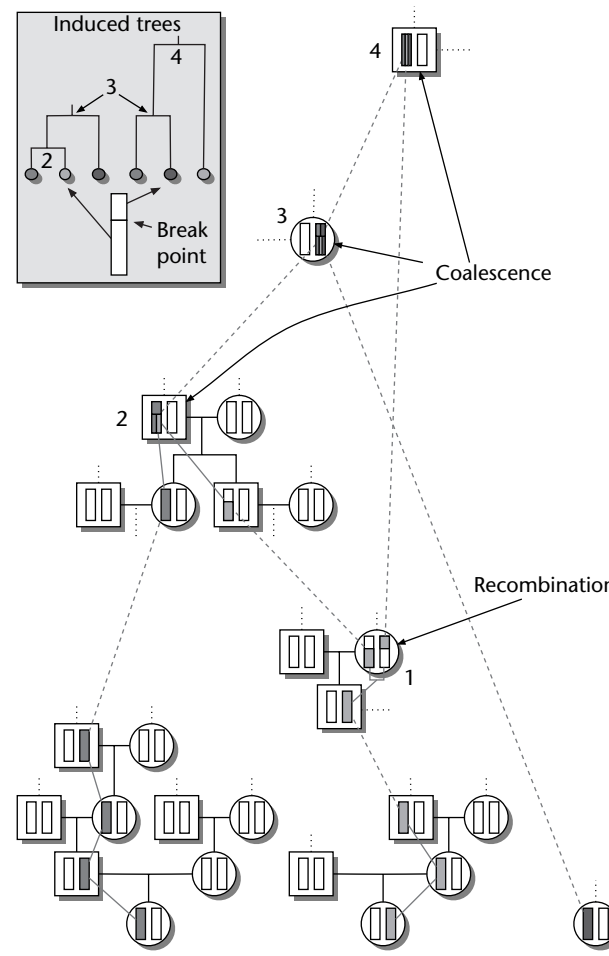

Figure 3 [*Figure is also reproduced in color section.*] Ancestral history of three chromosomal segments. See text for details.

We are not interested in the relationships among the individuals in this pedigree but rather the ancestral history of the three chromosomal segments highlighted at the bottom of the pedigree. There are four important events labeled in this history. At event 1, we observe a recombination event – the green segment now has two ancestors rather than one. At event 2, a coalescence event is shown. We see that the green segment and the purple segment share a common ancestral segment below the indicated recombination breakpoint. At event 3, we identify the segment that is ancestral to the entire blue and purple segments and the green segment below the recombination breakpoint. Finally, at event 4, we have identified the MRCA of all three segments. The three segments below the breakpoint have a common coalescent tree, and the three segments above the breakpoint have a different coalescent tree; these induced trees are shown in the inset in **Figure 3**. Because the two trees share some edges, they are not independent of each other.

The ancestral recombination graph (ARG) of a sample of n segments behaves in a similar way. Whenever a recombination event occurs, that ancestral segment has two ancestral segments; whenever a coalescence event occurs, the number of ancestral segments is reduced. In a large population, when there are k ancestral lines in the ARG, the relative rates of recombination (and thus an increase in the number of lines from k to $k+1$) and coalescence (and thus a decrease in the number of lines from k to $k-1$) are $k\rho/2$ and $k(k-1)/2$ respectively. Here, ρ is the scaled recombination rate; $\rho = 2Nr$, where r is the recombination fraction per segment per generation. The structure of the ARG for an arbitrary recombination rate across a segment is described in Griffiths and Marjoram (1997) and is illustrated for a sample of size 6 in **Figure 4** (adapted from Nordborg and Tavaré, 2002). Dot-dash lines indicate coalescence events and dotted lines recombination events. The chromosomal segment is described as a unit interval, and recombination breakpoints are indicated by locations in (0, 1); there are four such events in the figure, occurring at 0.14, 0.61, 0.93 and 0.98. By convention, a segment inherits the region to the left of a breakpoint from the left-hand ancestor,

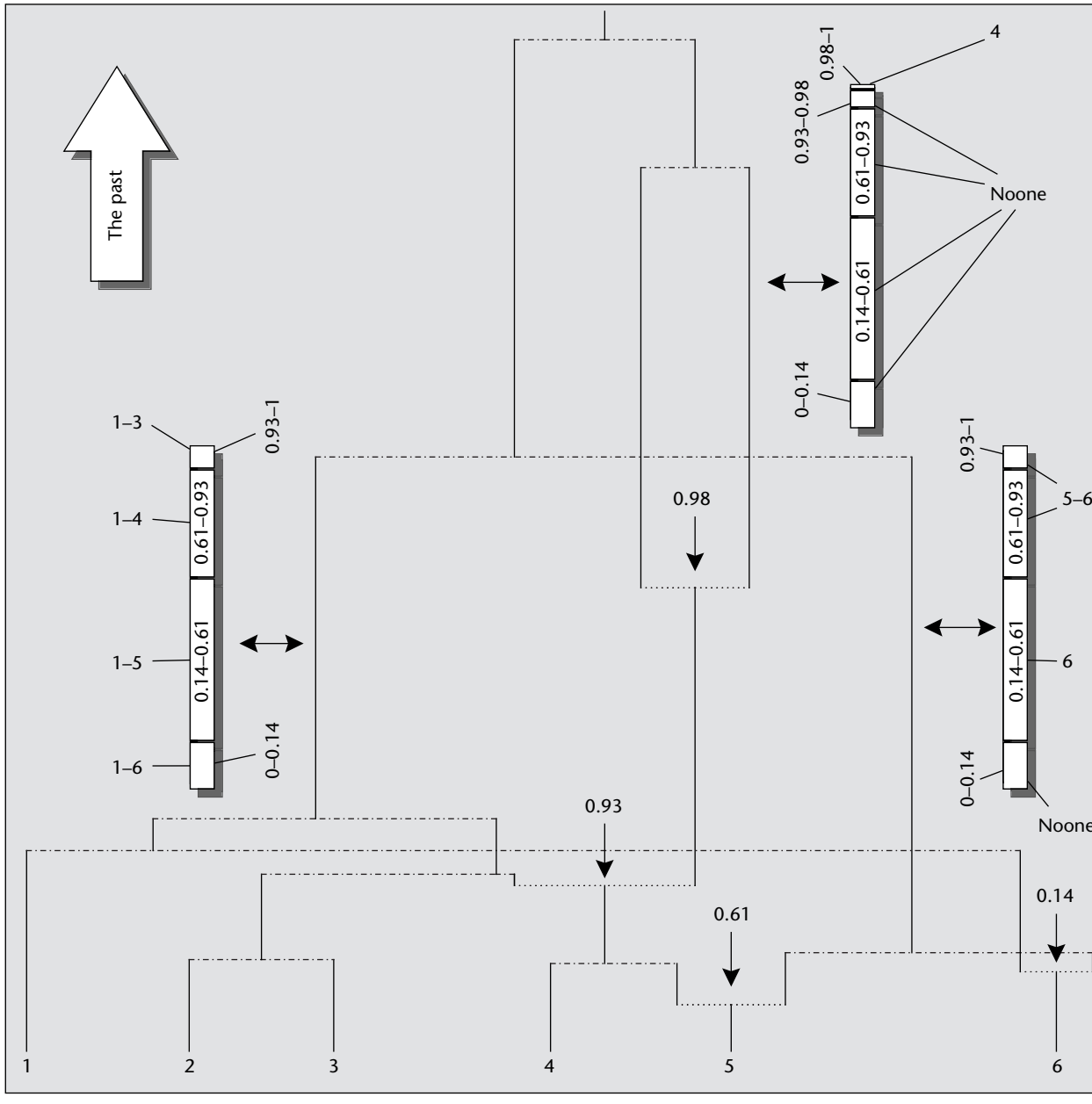

Figure 4 Ancestral recombination graph. Dot-dash lines indicate coalescence events; dotted lines indicate recombination events.

and the segment to the right from the right-hand ancestor. In **Figure 4**, three ancestral segments are drawn, showing which parts of each of them is ancestral to which members of the sample. As illustrated by the top segment, some genetic material in an ancestor need not be ancestral to any members of the sample.

Each point x in the segment has its own coalescent tree. These trees may be the same for different values of x; in particular, they are not independent of one another. The coalescent trees hidden in the ARG in **Figure 4** are shown in **Figure 5**.

Effects of mutation

Mutations are superimposed on the ARG just as in the case of no recombination, and their effects are also modeled in the same way. An obvious question concerns the relative sizes of ρ and θ. A direct estimate of θ can be obtained from sequence data; estimates on humans are typically about 10^{-3} per base pair (Przeworski *et al*., 2000). It is much harder to estimate ρ from polymorphism data, but the fact that $\rho/\theta = r/u$ and both r and u can be estimated directly provides another way to estimate ρ. For humans, a reasonable guess is that $\rho \approx \theta \approx 100$ for a 100-kb region.

Figure 6, from Rosenberg and Nordborg (2002), illustrates the interaction between mutation and recombination in a sample of size 6. It is based on simulation of an infinitely-many-sites model of a chromosomal segment of about 10 kb in length. The thin line in the figure shows the height of the coalescent tree at each point in the segment, and the locations of the mutations are shown as black dots. Note that only a few of the recombination events, shown as crosses, change the height of the coalescent tree. The thick line shows a moving average of the nucleotide diversity (the mean pairwise difference) across the region, the variation in that curve being due to randomness in the mutation process and the ARG.

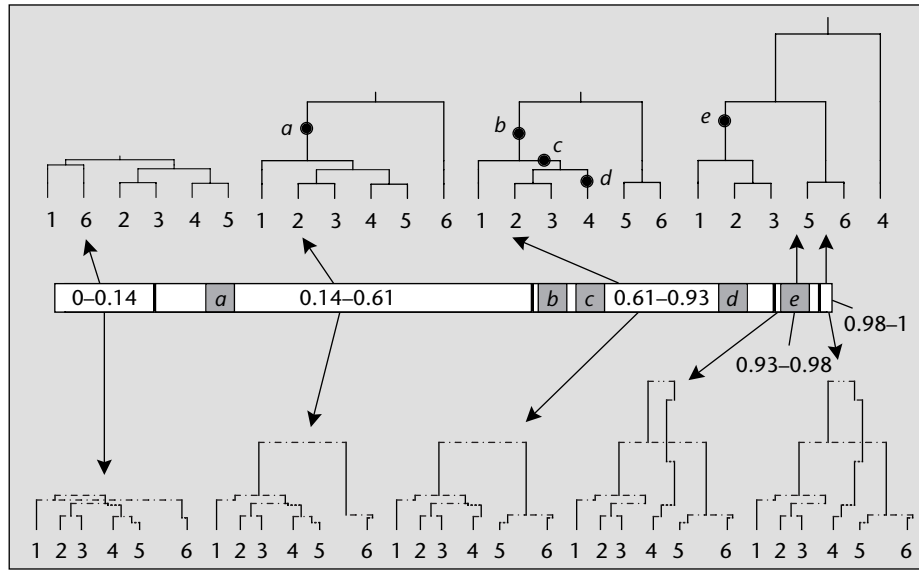

Figure 5 Coalescent trees in an ancestral recombination graph. The shaded points indicate mutations, which are represented as dots in each coalescent tree.

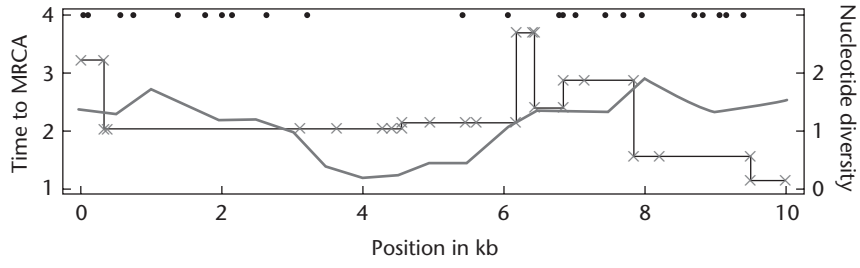

Figure 6 Nucleotide diversity and time to most recent common ancestor (MRCA) across a segment.

Linkage disequilibrium and haplotype sharing

Because the genealogical trees at different linked positions in a segment are not independent of one another, neither will be the allelic states of these loci: there will be *linkage disequilibrium* (LD) between the loci. LD is usually quantified by using various measures of association between pairs of loci. Consider two such loci, each of which has two possible alleles, and denote the relative frequency of the A_iB_j haplotype by $p(A_i, B_j)$, and let $p(A_i)$, $p(B_j)$ denote the relative frequency of each allele. Among the pairwise measures of LD are:

- D', the value of $D = p(A_1, B_1) - p(A_1)\, p(B_1)$, normalized to have values between -1 and 1 regardless of allele frequencies;
- r^2, the correlation in allelic state between the two loci as they occur in haplotypes;
- $d^2 = (p(B_1 \mid A_2) - p(B_1 \mid A_1))^2$, which measures the association between the alleles at (marker) locus B and the alleles at (disease) locus A.

These and other measures of LD are discussed further in Guo (1997), Hudson (2001) and Pritchard and Przeworski (2001).

Because of the history of recombination and mutation in a sample, pairwise LD is expected to be extremely variable. This is illustrated in **Figure 7**, adapted once more from Nordborg and Tavaré (2002). The horizontal axis, which represents chromosomal position, corresponds to roughly 100 kb. The plots illustrate the haplotype sharing and LD with respect to particular focal mutations. In the left column, a relatively low-frequency mutation ($5/50 = 10\%$) was chosen as focus, and in the right column, a relatively high-frequency one ($22/50 = 44\%$). The chromosomal positions of these mutations are indicated by the vertical lines. The top row of plots shows the extent of haplotype sharing with respect to the MRCA of the focal mutation among the 50 haplotypes. The horizontal lines indicate segments that descend from the MRCA of the focal mutation. A gray line indicates that the current haplotype also carries the focal mutation; black that it does not. Note that the gray segments necessarily overlap the position of the focal mutation. For clarity, segments that do not descend from the MRCA of the focal mutation are not shown at all, and haplotypes that do not carry segments descended from the MRCA of the focal mutation are therefore invisible. The second row of plots shows the behavior of LD as measured by d^2 for different choices of marker. In each plot, the horizontal position of a dot represents the chromosomal position of the marker, and the vertical position the value of the measure (on a zero-to-one scale).

Because of interest in mapping disease-susceptibility genes, the extent of LD across the human genome has been much debated. What is clear is that while there is a relationship between LD measures and distance, the inherent variability in LD makes this relationship hard to infer. In particular, it is difficult to compare studies that use different measures of pairwise LD as these measures can differ dramatically in their estimates of the range of LD. For reviews of these issues in relation to mapping, see, for example, Weiss and Clark (2002), Nordborg and Tavaré (2002) and Ardlie *et al*. (2002).

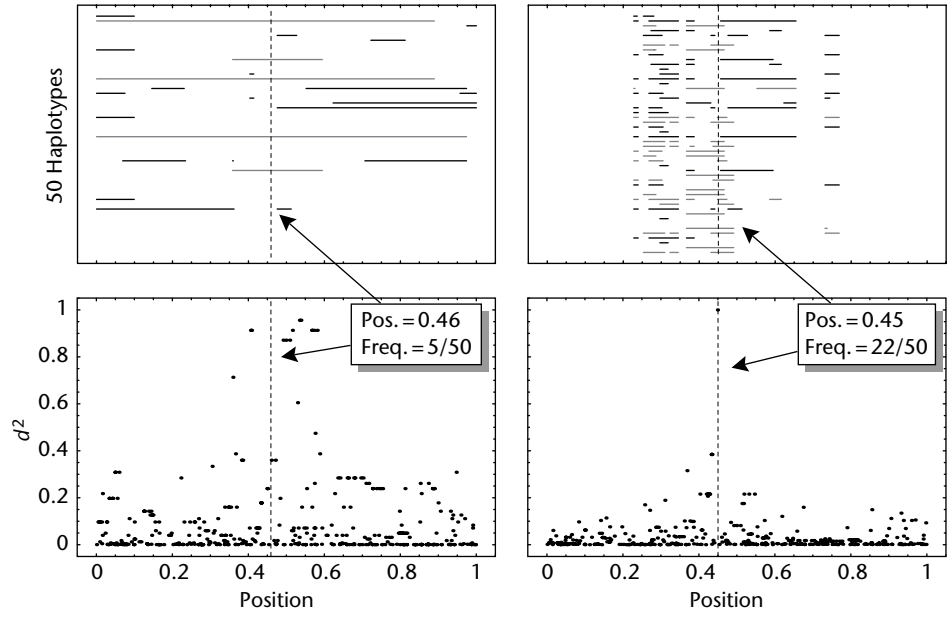

Figure 7 Decay of haplotype sharing.

Selection

The previous discussion has focused on neutral genes, in which the effects of mutation could be superimposed on top of the underlying coalescent genealogy. When selection is acting at some loci, this separation is no longer possible, and the analysis is rather more complicated. Two basic approaches have emerged. In the first, the evolution of the selected loci is modeled forward in time, and then the neutral loci are studied by coalescent methods (Kaplan *et al*., 1988, 1989). In the second a genealogical process known as the *ancestral selection graph*, the analog of the neutral coalescent, is developed (Neuhauser and Krone, 1997; Krone and Neuhauser, 1997). Neuhauser (2001) and Nordborg (2001) contain reviews on the subject. Methods for simulating selected genealogies are an important current area of research (Slatkin, 2001; Slade, 2001; Fearnhead, 2001). Such simulations can be used to explore the consequences of different selection mechanisms on the pattern of variation observed in data. Methods for detecting selection in sequence data are reviewed in Kreitman (2000).

Inference in the Coalescent

Statistical theory for coalescent models has developed in two ways. Both employ simulation, albeit in rather different ways. The first is a *top-down* approach: multiple data sets are simulated from a given coalescent model, and summary statistics of these simulated data sets are compared to observed data. The second method is a *bottom-up* approach that uses explicit model-based likelihood calculations. Here, simulation methods such as Markov chain Monte Carlo (MCMC) are used for computational evaluation of distributions of interest. We now describe these approaches in more detail.

Forward simulation

The simple stochastic structure of coalescent models makes the simulation of typical (neutral) data sets a relatively easy task. Many different scenarios, involving for example migration, subdivision and recombination, can be explored this way; Hudson (1990) describes several examples. One statistical use for this type of simulation is to investigate the distribution of test statistics under neutrality, a classical example being provided by Tajima's D statistic, commonly used to detect selection in DNA sequence data (Tajima, 1989). Simulation can also be used to examine the robustness of such null distributions to changes in demography.

Backward simulation

The forward simulations described in the section above generate *typical* data sets. For inference concerning a *particular* data set, it is necessary to explore the space of parameters and ancestral histories consistent with that data set. Methods for doing this are considerably more complicated than forward methods, but they do provide a rigorous framework for estimation and inference in the coalescent. The early research in this field concentrated on computational methods for estimating the mutation parameter θ from a sample of sequences. Griffiths and Tavaré (1994a, b) used a Markov chain method to approximate the likelihood of a sample on a grid of θ points, while Kuhner *et al*. (1995) pioneered the use of Markov chain Monte Carlo (MCMC) methods for this purpose. Both approaches have been refined and extended. Griffiths and Marjoram (1996) describe an approach to maximum likelihood estimation of recombination rates for the infinite-sites model with recombination. Felsenstein *et al*. (1999) identified the Griffiths-Tavaré method as a version of importance sampling, which opened up the method to the development of better sampling schemes (Stephens and Donnelly, 2000), and further improvements using sequential importance sampling (cf. chap. 2.7 of Liu (2001). The MCMC methods have been developed for a number of models, including migration (Beerli and Felsenstein, 1999, 2001) and recombination (Kuhner *et al*., 2000). Stephens (2001) provides a comprehensive review of many of these methods and describes some of the software that has been implemented.

A Bayesian perspective

Described here is a relatively new approach to inference in the coalescent that can be used in both Bayesian and likelihood-based approaches to inference.

Denote the ancestral tree (or graph) of the sample by Λ, the parameters of interest (population sizes, recombination rates, etc.) by μ and the data (typically DNA sequences) obtained from the sample by $\mathfrak{D}$. The Bayesian approach to inference in the coalescent proceeds by calculating the posterior distribution $f(\Lambda, \mu, |\mathfrak{D})$. Note that

$$f(\Lambda, \mu \,|\, \mathfrak{D}) \propto \mathbb{P}(\mathfrak{D} \,|\, \Lambda, \mu)\pi(\Lambda, \mu)$$

where $\pi(\cdot,\cdot)$ denotes the prior distribution of (Λ,μ). The obvious computational approach to simulating observations from f proceeds via the rejection method (Tavaré *et al*., 1997):

(R1) Simulate (Λ, μ) from π.
(R2) Calculate $p = \mathbb{P}(\mathfrak{D}|\Lambda, \mu)$.
(R3) Keep (Λ, μ) with probability p

It is well known that the accepted observations are independent and have the required distribution f. Note that π incorporates information about the coalescent model and prior beliefs about the parameters in that model.

In practice, this approach can fail in several ways. For example, the acceptance probability p can be so small that observations are rarely accepted. This problem has also been addressed by MCMC methods. usually as a version of the Metropolis–Hastings algorithm. Denoting the current state of the process by $G=(\Lambda, \mu)$, this algorithm proceeds as follows:

(MH1) Output the current value of G.
(MH2) Propose $G'=(\Lambda', \mu')$ according to a kernel $Q(G \to G')$.
(MH3) Compute the Hastings ratio:

$$h=\frac{f(G' \mid \mathfrak{D})Q(G' \to G)}{f(G \mid \mathfrak{D})Q(G \to G')}$$

(MH4) Accept the new state G' with probability $\min(1, h)$, otherwise stay at G.

After a burn-in period, this algorithm produces (under some regularity conditions on the moves determined by Q) correlated samples having (approximately) the distribution f. These samples can be used to make inferences about the posterior distribution of parameters and statistics of interest. Examples include the mutation rate θ, the time to the most recent common ancestor, ages of a particular event in the sample or population growth rates. Different implementations are often used for different mutation mechanisms; for example, Wilson and Balding (1998) treat microsatellite data and Markovtsova *et al.* (2000a) treat DNA sequence data. The Bayesian approach also provides a natural way to examine goodness-of-fit of models to data via the posterior predictive distribution (Markovtsova *et al.*, 2000a).

It is worth noting that the Bayesian approach often provides an efficient computational approach to maximum likelihood estimation for μ. As eqn [1] shows, $\mathbb{P}(\mathfrak{D} \mid \mu) \propto f(\mu \mid \mathfrak{D})/\pi(\mu)$, so the likelihood can be estimated, up to a scale parameter, from the prior and the posterior (see Stephens and Donnelly, 2000; Markovtsova *et al.*, 2000b, for example).

Approximate Bayesian methods

MCMC methods are quite complicated to implement and very difficult to check carefully. In particular, there are well-known difficulties in deciding when the chain has 'reached stationarity' (Gilks *et al.*, 1996). Furthermore, small changes in the problem to be addressed (e.g. incorporating population expansion into a model that currently lacks it) can result in drastic changes in the algorithm. The Metroplis–Hastings algorithm defined above also uses the values of the likelihood $\mathbb{P}(\mathfrak{D} \mid \Lambda, \mu)$; in calculating the Hastings ratio in (MH3) above, ratios of likelihoods are required. Computing such likelihoods, for example by a peeling algorithm for DNA sequence data, is extremely time consuming. In some cases, computation of this likelihood is impossible, numerically or otherwise. An interesting alternative approach, based on approximate Bayesian computation, has been introduced by Weiss and von Haeseler (1998) and Pritchard *et al.* (1999).

This method takes advantage of the relative ease with which coalescent-like processes can be simulated and simultaneously exploits the advantages of the rejection method. The method can be summarized as follows:

(AB1) Simulate (Λ, μ) from π.
(AB2) Simulate a data set $\mathfrak{D}'$ from the coalescent model *with these parameters*.
(AB3) For a given distance measure $d(\mathfrak{D}, \mathfrak{D}')$ on the space of possible data sets, accept (Λ, μ) if $d(\mathfrak{D}, \mathfrak{D}') < \epsilon$, for some cutoff value ϵ.

The accepted observations have a distribution of the form $f^*(\Lambda, \mu \mid d(\mathfrak{D}, \mathfrak{D}') < \epsilon)$. If $\epsilon = 0$, the accepted values have the same summary statistics as the original data, as determined by the distance d. In practice, the art in the method comes from judicious choice of d and ϵ: if $\epsilon = 0$, very few observations may be accepted, whereas if $\epsilon = \infty$ the simulated values come from the prior π. For sequence data, it is impractical to expect every feature of the simulated sequences to match those of the given data, but sufficient information might be captured by requiring, for example, that the number of segregating sites and the number of distinct haplotypes be similar. Wall (2000) exploited a version of this approach in his study of estimators of the recombination rate ρ. There has been as yet no formal study of methods for choosing useful summary statistics. The hope is that experience will produce a reasonably flexible, easy-to-implement and fast method of (usefully) approximate inference.

Ancestral inference

Under the coalescent model, the height of a coalescent tree is an unobserved random variable; it is therefore natural to report its conditional distribution given the data $\mathfrak{D}$ (Tavaré *et al.*, 1997). Such times are of interest, for example, in the study of human evolution (Pritchard *et al.*, 1999). The Bayesian methods also generate observations from the posterior distribution of the coalescent tree itself, in particular its height. Such observations can be used to study the required posterior distribution.

While much of the focus of the inference methods discussed above has been on estimates or posterior distributions of parameters of the model, the same methods can be used to study the age of a unique event polymorphisms, those mutations that arise just once in the history of a population. There have been numerous approaches to this problem, summarized in Slatkin and Rannala (2000). Some explicit coalescent theory is available to describe the age of an allele now known to have a given frequency in a sample (Griffiths and Tavaré, 1998; Wiuf and Donnelly, 1999). When further information is given about the variability of the region around the unique event polymorphism, more computationally intensive methods are required; see Markovtsova *et al*. (2000b) for an example.

Discussion

This brief article has focused on the structure of the coalescent and its relatives, and on the basic methods of inference for such processes. The field is currently undergoing a dramatic expansion as its usefulness for interpreting genomic variation data is more fully appreciated. There are still many challenging problems, including inference for models with selection and the development of robust methods of inference for haplotype data, in particular for case–control and similar sampling designs in use in human genetics. It seems likely that coalescent theory will shed useful light on the haplotype block structure in the human genome (Daly *et al*., 2001; Johnson *et al*., 2001; Patil *et al*., 2001; Gabriel *et al*., 2002). Methods that use the 'genomic controls' provided by chromosome-wide data in testing for signs of selection at particular loci are to be preferred to those based on models that may not apply well enough to the data at hand. However, this remains a complicated issue that is exacerbated by the multiple comparison problem. This is but one of the data analysis problems that are arising from genome-wide data.

Acknowledgement

The author was supported in part by NIH grant GM 58897.

See also

Diffusion Theory
Fixation Probabilities and Times
Population Genetics

References

Ardlie KG, Kruglyak L and Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics* **3**: 299–309.

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

Beerli P and Felsenstein J (2001) Maximum-likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 4563–4568.

Cannings C (1974) The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. *Advances in Applied Probability* **6**: 260–290.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ and Lander ES (2001) High-resolution haplotype structure in the human genome. *Nature Genetics* **29**: 229–232.

Ewens WJ (1979) *Mathematical Population Genetics*. Berlin, Germany: Springer-Verlag.

Ewens WJ (1990) Population genetics theory – the past and the future. In: Lessard S (ed.) *Mathematical and Statistical Developments of Evolutionary Theory*, pp. 177–227. Dordrecht, Holland: Kluwer Academic.

Fearnhead P (2001) Perfect simulation from population genetic models with selection. *Theoretical Population Biology* **59**: 263–279.

Felsenstein J, Kuhner MK, Yamato J and Beerli P (1999) In: Seillier-Moisiewitsch F (ed.) *Statistics in Molecular Biology and Genetics*, pp. 163–185. Lecture Notes Monograph Series, vol. 33. Hayward, CA: IMS.

Gabriel SB, Schaffner SF, Nguyen H, *et al*. (2002) The structure of haplotype blocks in the human genome. *Science* **296**: 2225–2229.

Gilks WR, Richardson S and Spiegelhalter DJ (1996) *Markov chain Monte Carlo in Practice*. London, UK: Chapman & Hall.

Griffiths RC (1991) The two-locus ancestral graph. In: Basawa IV and Taylor RL (eds.) *Selected Proceedings of the Sheffied Symposium on Applied Probability*, pp. 100–117. Lecture Notes-Monograph Series, vol. 18. Hayward, CA: IMS.

Griffiths RC and Marjoram P (1996) Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology* **3**: 479–502.

Griffiths RC and Marjoram P (1997) An ancestral recombination graph. In: Donnelly P and Tavaré S (eds.) *Progress in Population Genetics and Human Evolution*, pp. 257–270. New York, NY: Springer-Verlag.

Griffiths RC and Tavaré S (1994a) Ancestral inference in population genetics. *Statistical Science* **9**: 307–319.

Griffiths RC and Tavaré S (1994b) Simulating probability distributions in the coalescent. *Theoretical Population Biology* **46**: 131–159.

Griffiths RC and Tavaré S (1998) The age of a mutant in a general coalescent tree. *Stochastic Models* **14**: 273–295.

Guo S-W (1997) Linkage disequilibrium measures for fine-scale mapping: a comparison. *Human Heredity* **47**: 301–314.

Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23**: 183–201.

Hudson RR (1990) Gene genealogies and the coalescent process. In: Futuyma D and Antonovics J (eds.) *Oxford Surveys in Evolutionary Biology*, vol. 7, pp. 1–43. Oxford, UK: Oxford University Press.

Hudson RR (2001) Linkage disequilibrium and recombination. In: Balding DJ, Bishop MJ and Cannings C (eds.) *Handbook of Statistical Genetics*, pp. 309–324. Chichester, UK: Wiley.

Johnson GCL, Esposito L, Barratt BJ, *et al*. (2001) Haplotype tagging for the identification of common disease genes. *Nature Genetics* **29**: 233–237.

Kaplan NL and Hudson RR (1985) The use of sample genealogies for studying a selectively neutral *m*-loci model with recombination. *Theoretical Population Biology* **28**: 382–396.

Kaplan NL, Darden T and Hudson RR (1988) The coalescent process in models with selection. *Genetics* **120**: 819–829.

Kaplan NL, Hudson RR and Langley CH (1989) The 'hitch-hiking' effect revisited. *Genetics* **123**: 887–899.

Kingman JFC (1982a) On the genealogy of large populations. In: Gani J and Hannan EJ (eds.) *Essays in Statistical Science: Papers in Honour of P.A.P. Moran*, pp. 27–43. *Journal of Applied Probability* **19A**.

Kingman JFC (1982b) Exchangeability and the evolution of large populations. In: Koch G and Spizzichino F (eds.) *Exchangeability in Probability and Statistics*, pp. 97–112. Amsterdam, Holland: North-Holland Publishing.

Kreitman M (2000) Methods to detect selection in populations with applications to the human. *Annual Review of Genomics and Human Genetics* **1**: 539–559.

Krone SM and Neuhauser C (1997) Ancestral processes with selection. *Theoretical Population Biology* **51**: 210–237.

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

Kuhner MK, Yamato J and Felsenstein J (2000) Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**: 1393–1401.

Liu J (2001) *Monte Carlo Strategies in Scientific Computing*. New York, NY: Springer-Verlag.

Markovtsova L, Marjoram P and Tavaré S (2000a) The effects of rate variation on ancestral inference in the coalescent. *Genetics* **156**: 1427–1436.

Markovtsova L, Marjoram P and Tavaré S (2000b) The age of a unique event polymorphism. *Genetics* **156**: 401–409.

Möhle M (1998) Robustness results for the coalescent. *Journal of Applied Probability* **35**: 438–447.

Möhle M and Sagitov S (2001) A classification of coalescent processes for haploid exchangeable population models. *Annals of Probability* **29**: 1547–1562.

Neuhauser C (2001) Mathematical models in population genetics. In: Balding DJ, Bishop MJ and Cannings C (eds.) *Handbook of Statistical Genetics*, pp. 153–177. New York, NY: Wiley.

Neuhauser C and Krone SM (1997) The genealogy of samples in models with selection. *Genetics* **145**: 519–534.

Nordborg M (2001) Coalescent theory. In: Balding DJ, Bishop MJ and Cannings C (eds.) *Genetics*, pp. 179–212. Chichester, UK: Wiley.

Nordborg M and Tavaré S (2002) Linkage disequilibrium: what history has to tell us. *Trends in Genetics* **18**: 83–90.

Patil N, Berno AJ, Hinds DA, *et al*. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719–1723.

Pritchard JK and Przeworski M (2001) Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics* **69**: 1–14.

Pritchard JK, Seielstad MT, Perez-Lezaun A and Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution* **16**: 1791–1798.

Provine WB (2001) *The Origins of Theoretical Population Genetics*, 2nd edn. Chicago, IL: University of Chicago Press.

Przeworski M, Hudson RR and Di Rienzo A (2000) Adjusting the focus on human variation. *Trends in Genetics* **16**: 296–302.

Rosenberg NA and Nordborg M (2002) Genealogical trees, coalescent theory, and the analysis of genetic polymorphisms. *Nature Reviews Genetics* **3**: 380–390.

Slade PF (2001) Simulation of 'hitch-hiking' genealogies. *Journal of Mathematical Biology* **42**: 41–70.

Slatkin M (2001) Simulating genealogies of selected alleles in a population of variable size. *Genetic Research* **78**: 49–57.

Slatkin M and Rannala B (2000) Estimating allele age. *Annual Review of Genomics and Human Genetics* **1**: 225–249.

Stephens JC, Schneider JA, Tanguay DA, *et al*. (2001) Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**: 489–493.

Stephens M (2001) Inference under the coalescent. In: Balding DJ, Bishop MJ and Cannings C (eds.) *Handbook of Statistical Genetics*, pp. 213–238. Chichester, UK: Wiley.

Stephens M and Donnelly P (2000) Inference in molecular population genetics. *Journal of the Royal Statistical Society, B* **62**: 605–655.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**: 437–460.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.

Tavaré S (1984) Line-of-descent and genealogical processes, and their applications in population genetic models. *Theoretical Population Biology* **26**: 119–164.

Tavaré S, Balding DJ, Griffiths RC and Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics* **145**: 505–518.

Thorne JL, Kishino H and Felsenstein J (1992) Inching towards reality: an improved likelihood model of sequence evolution. *Journal of Molecular Evolution* **34**: 3–16.

Wall JD (2000) A comparison of estimators of the population recombination rate. *Molecular Biology and Evolution* **17**: 156–163.

Weiss KM and Clark AG (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends In Genetics* **18**: 19–24.

Weiss G and von Haeseler A (1998) Inference of population history using a likelihood approach. *Genetics* **149**: 1539–1546.

Wilson IJ and Balding DJ (1998) Genealogical inference from microsatellite data. *Genetics* **150**: 499–510.

Wiuf C and Donnelly P (1999) Conditional genealogies and the age of a neutral mutant. *Theoretical Population Biology* **56**: 183–201.

Further Reading

Beaumont MA, Zhang W and Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics* **162**: 2025–2035.

Dawson E, Abecasis GR, Bumpstead S, *et al*. (2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**: 544–548.

Fearnhead P and Donnelly P (2002) Approximate likelihood methods for estimating local recombination rates. *Journal of the Royal Statistical Society Series B*, **64**: 657–680.

Kong A, Gudbjartsson DF, Sainz J, *et al.* (2002) A high-resolution recombination map of the human genome. *Nature Genetics* **31**: 241–247.

Morris AP, Whittaker JC and Balding DJ (2002) Fine scale mapping of disease loci via shattered coalescent modelling of genealogies. *American Journal of Human Genetics*, **70**: 686–707.

Reich DE, Schaffner SF, Daly MJ, *et al*. (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genetics*, **32**: 135–142.

Tang H, Siegmund DO, Shen P, Oefner PJ and Feldman MW (2002) Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. *Genetics* **161**: 447–459.