Journal of Heredity 2004:95(5):397–405
doi:10.1093/jhered/esh062

Recent Trends in Population Genetics: More Data! More Math! Simple Models?

J. WAKELEY

From the Department of Organismic and Evolutionary Biology, Harvard University, 2102 Biological Laboratories, 16 Divinity Ave., Cambridge, MA 02138. I thank Kent Holsinger for the invitation to participate in the AGA centenary celebration and for helpful comments on the manuscript. This work was supported by a Presidential Early Career Award for Scientists and Engineers (DEB-0133760) from the National Science Foundation.

Address correspondence to John Wakeley at the address above, or e-mail: wakeley@fas.harvard.edu.

Abstract

Recent developments in population genetics are reviewed and placed in a historical context. Current and future challenges, both in computational methodology and in analytical theory, are to develop models and techniques to extract the most information possible from multilocus DNA datasets. As an example of the theoretical issues, five limiting forms of the island model of population subdivision with migration are presented in a unified framework. These approximations illustrate the interplay between migration and drift in structuring gene genealogies, and some of them make connections between the fairly complicated island-model genealogical process and the much simpler, unstructured neutral coalescent process which underlies most inferential techniques in population genetics.

The field of population genetics has undergone remarkable changes in the past few decades. This has been driven mostly by the development of DNA sequencing technologies, which now make gathering large quantities of the most direct kind of genetic data easy and affordable. Theoretical models and computational techniques appropriate to handle these data are still in development, and there is great need for further work. This article gives a short history of the field in relation to these developments and outlines some of the mathematical issues relevant to the study of gene genealogies of samples from demographically complicated populations. These sorts of analyses, which sometimes yield surprisingly simple results, are illustrated for genetic ancestries of samples of size two in Wright's (1931) island model of population structure, but the conclusions are limited neither to such small samples nor to such simple population structures.

Theoretical Population Genetics History

The story of the emergence of theoretical population genetics, out of a tension between biometricians and Mendelians, has been told eloquently by Provine (1971). In relation to the current state of the field, it is interesting to note that even the first population genetics theory was data driven. Fisher (1918), in an article often taken to represent the birth of the field, used mathematics to show that two apparently conflicting sets of available data were actually in perfect harmony. In particular, Fisher (1918) demonstrated that measured correlations between relatives, which were the focus of biometricians' studies, could be explained by the contributions of a large number of Mendelian factors (now, polymorphic loci) each of small effect. It was in that same article that Fisher introduced variance and covariance as the most natural and convenient measures of dispersion and correlation, showing, for example, that it is much easier to separate out contributions to the variance than it is to decompose the standard deviation, which was the favored measure of the biometricians. It was in this and subsequent articles that Fisher developed the method of analysis of variance (ANOVA), which became a mainstay of statistical data analysis.

The early works of Fisher (1930), Wright (1931), and Haldane (1932) built the foundation of theoretical population genetics and established many of the fundamental results still quoted today. During the period from about 1940 to the mid-1960s, these and other authors produced many more detailed mathematical results about the evolutionary process and about the maintenance of genetic variation within populations. In addition, this period saw the extension of the field into an even more sophisticated mathematical realm by such notables as Malécot (1948) and Kimura (1955a,b). This work proceeded without the benefit of direct genetic data (Lewontin 1974), but can now be seen to form the basis of the next data-driven advancement, which came with the introduction of gel electrophoresis to population genetics by Harris (1966) and Lewontin and Hubby (1966).

Again, the availability of data, in this case measurements of allozyme variability within and among populations, spurred the development of new theory. Ewens (1972) proposed a new statistical distribution that predicted patterns of selectively neutral allozyme variation in a sample from a large population. The introduction of the "Ewens sampling formula" marks the beginning of a shift in perspective from a prospective view of classical population genetics to a new, retrospective view which was soon embodied by Kingman's (1982a,b) coalescent; see Ewens (1990) for a discussion of these developments. Whereas the classical approach used forward-time analyses to make predictions about genetic variation in a population and required a separate theory of sampling, this new work took a backwards-time approach to generate directly, predictions about genetic variation in a sample. Thus the retrospective approach has always been closely tied to samples and to inference. One early example is Watterson (1977), who noted that the distribution of allele frequencies in a sample could contain information about the action of natural selection and proposed a test for selection based on deviations from the Ewens distribution.

Kingman's Coalescent

Ewens developed his sampling formula using the notion of identity by descent, which had been introduced by Malécot (1946), and under the assumption of infinite alleles mutation (Kimura and Crow 1964; Malécot 1946). This prompted a series of works by Watterson (1976a,b), Griffiths (1979,1980), and others, describing the diffusion approximation (thus building on Kimura's work) for the neutral, infinite alleles model. Because alleles in the infinite alleles model are always related in the genealogical sense, this work was instrumental in the next major development in population genetics (Kingman 2000), which was the introduction of the coalescent process by Hudson (1983a,b), Kingman (1982a,b), and Tajima (1983). Another precurser to the coalescent process was Watterson (1975), in which predictions about levels of sequence variation in a sample were made, using genealogical ideas, under the assumption of infinite sites mutation (Kimura 1969) without recombination. Under the infinite sites model or others appropriate for DNA, the coalescent is well suited for the analysis of sequence data. It is not just a coincidence that the introduction of the coalescent coincided with the first application of DNA sequencing technology to the problem of measuring genetic variation (Kreitman 1983).

Donnelly and Tavaré (1995), Hudson (1990), and Nordborg (2001) provide reviews of coalescent theory. Briefly, under the assumption of selective neutrality it is possible to model just the history of a sample, that is, without regard to the rest of the population. Selection can be accommodated easily if it is strong (Kaplan et al. 1988, 1989), while coalescent models of weak selection (Krone and Neuhauser 1997; Neuhauser and Krone 1997) are more complicated. The coalescent, as it is typically presented in population genetics, makes all the usual assumptions of the Wright-Fisher model of a population (Fisher 1930; Wright 1931). In addition to selective neutrality, it is assumed that the population is of constant size and is not structured in any way (by geography, gender, age, or nonrandom mating). The latter makes the members of a sample, or the ancestral lineages of a sample as they are followed back in time, exchangeable in the statistical sense (Aldous 1985; Kingman 1982b), which means that they are not distinguished by any properties that affect rates of coalescence.

When time is measured in units of $N_e = N/\sigma^2$ generations, where N is the population size and σ^2 is the variance in offspring numbers among members of the population, and the effective size N_e is large, then the rate of coalescence is equal to one for every pair of sample lineages (Kingman 1982a,c). Further, every coalescent event involves just two lineages, so the history of the sample of size n back to the most recent common ancestor includes exactly $n - 1$ coalescent events. The times T_i between coalescent events are distributed exponentially and depend on the number i of lineages present during each interval:

$$f_{T_i}(t) = \binom{i}{2} e^{-\binom{i}{2}t}, \qquad (1)$$

where $\binom{i}{2} = i(i-1)/2$ is the number of possible pairs of i lineages. In the special case of a sample of size two, the time to the most recent common ancestor is exponentially distributed with rate equal to one (i.e., putting in $i = 2$ above). The exchangeability of lineages is reflected in the fact that when a coalescent event occurs among the members of a sample, every pair of lineages is equally likely to be the pair that coalesces.

Formally, Equation (1) is obtained for a fixed sample size n in many exchangeable population models (Cannings 1974), as the population size N tends to infinity and time is measured appropriately (in units of $N_e = N/\sigma^2$ generations). In the limit $N \to \infty$, the possibility of multiple coalescent events in a single generation becomes negligible and the discrete-time process of genetic ancestry is replaced by the continuous-time process embodied in Equation (1). The resulting model is used as an approximation to the ancestral process for samples and populations in which the sample size is much less than the population size ($n \ll N$).

Recent Trends in Population Genetics

The past few years have seen an explosion of DNA sequencing and other genotyping technologies as a result of the genome projects of humans and other organisms. Technical improvements, such as the use of robotics, have found their way into most universities and streamlined the gathering of relatively large genetic datasets even in nonmodel organisms. In particular, it is now common to see analyses of multiple genetic loci, whereas 20 years ago it was a major challenge to obtain sequence data from a single locus. This is of fundamental importance to the field of population genetics because we can expect to uncover from multiple loci both genome-wide patterns and locus-specific effects. Population structure is an example of a phenomenon

Table 1. Theoretical predictions and observed counts of polymorphic sites for samples of size two at 11,027 human genetic loci.

No. of SNPs	Poisson	Coalescent	Observed
0	8256 ± 52	8767 ± 50	8796 ± 43
1	3040 ± 49	2332 ± 46	2247 ± 44
2	617 ± 24	663 ± 26	668 ± 24
3	99 ± 9	200 ± 15	214 ± 14
4	16 ± 4	66 ± 9	102 ± 10

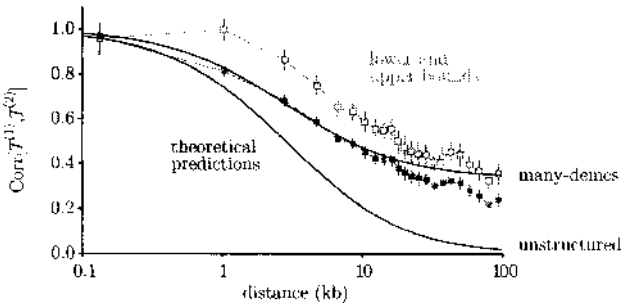

Figure 1. Estimated correlations of genealogical tree lengths at pairs of loci separated by different distances in the human genome; redrawn from Figure 2a in Wakeley and Lessard (2003), which corresponds to Figure 5a of Reich et al. (2002). Theoretical prediction is for an unstructured population of size $N_e = 10^4$ and a recombination rate per base pair per meiosis of 1.3×10^{-8}. See Reich et al. (2002) and Wakeley and Lessard (2003) for details.

that affects loci across the genome in a similar manner, while natural selection is an example of processes that can affect single loci. It may be impossible to distentangle the forces that have produced and maintained variation at a single locus without having a genomewide picture of variation because single loci represent just one realization of the stochastic and multifactorial process of descent within populations.

At present, the datasets with the largest number of loci are from humans and model organisms such as *Arabidopsis*, *Drosophila*, and mouse. An examination of some observations from human population genetics helps to illustrate the future hopes and challenges for the field. For example, Table 1, which is redrawn from Table 3 of the International SNP Map Working Group (2001), shows theoretical predictions and observed counts of polymorphic sites for samples of size two at 11,027 human genetic loci spread more or less randomly throughout the genome. The table shows that a simple Poisson prediction, which would hold if there was no variation in coalescent times among loci, fits the data very poorly, and that predictions from the standard coalescent provide a much better fit. However, the fit of the coalescent prediction is also poor ($\chi^2 = 23.85$; $P < .01$), indicating that one or more of the assumptions of the standard coalescent model does not hold for humans. Other analyses of multilple loci similarly conclude that simple models cannot explain the data (Pluzhnikov et al. 2002; Przeworski et al. 2000).

Thus there appears in multilocus data from humans to be information about other processes—that is, migration, changes in population size, and/or natural selection—than those modeled in the standard coalescent. This is of course not surprising given the dynamic history of humans (Takahata 1995; Harpending et al. 1998; Hawks et al. 2000), but rather offers the hope that inferences might be made about some of these more complicated and interesting phenomena. Another example comes from a more detailed study by Reich et al. (2002) of a similar but much larger dataset from humans. Reich et al. (2002) measured correlations in genealogical tree lengths (or coalescent times) between pairs of loci separated by different distances along the genome. One of the results of their analyses is depicted in Figure 1. Correlations in genealogical tree lengths are expected to decline with the distance between loci due to recombination, and Reich et al. (2002) showed that a prediction for this decline based on the standard coalescent with recombination (lower black curve) could not explain the long-range correlations in the human genome. Interestingly, a prediction from one of the models of population structure with migration considered below (many-demes model: upper black curve) may be at least a partial explanation for these correlations (Wakeley and Lessard 2003).

Multilocus data such as those presented in Table 1 and Figure 1 motivate current work both on theoretical models and statistical techniques. Broadly put, the aim is to develop models that include all the relevant processes and to produce a suite of inferential methods that use multilocus data to tease apart the effects of multiple forces acting simultaneously. Stephens (2001) and Tavaré (2004) review trends in the development of statistical techniques. Briefly, these center around the problem of computing the likelihood $P(\text{data}|\text{model})$, which is the probability of the observed data under a model with specified values of all parameters. A first step is to condition on the underlying genealogy, since $P(\text{data}|\text{genealogy,model})$ is usually easy to compute. Then, because it is nearly impossible to "integrate" over genealogies analytically, these are generated randomly using simulations and $P(\text{data}|\text{genealogy,model})$ is averaged over many genealogies. Methods differ in how genealogies are produced, specifically in how the information in the data is used to inform the choice of genealogies, and in whether inferences are based on the likelihood or computation of $P(\text{data}|\text{model})$ is imbedded in a bayesian method of inference. The inclusion of additional factors, such as migration and recombination, adds to the computational complexity of the problem because it expands the space of genealogies and because inferences must then be made in a multidimensional parameter space.

The development of theoretical models that can aid in understanding complicated demographic histories and provide a basis for methods of statistical inference has been another major aim of recent work. In addition to natural selection, mentioned above, the genealogical models have been extended to include changes in population size (Kingman 1982a; Slatkin and Hudson 1991), recombination

(Hudson 1983a; Kaplan and Hudson 1985), migration (see below), and sometimes several of these factors at once (Kaplan et al. 1991). One of the important roles of analytical work is to identify cases in which the structure of complicated, multiparameter models reduces to something simpler. When this is possible, it can lead to greater understanding of the interplay of processes affecting data as well as to more efficient computational techniques. Results of this sort come from studying the limiting behavior of a model as one (or more) of the parameters becomes either large or small. The question is then whether any of these simpler models are appropriate for modeling the history of a particular species. To illustrate the techniques and give an example of such results, the following section describes five mathematical limits of a commonly employed model of population subdivision with migration.

Coalescence in the Island Model and Simplifications

Wright's (1931) island model of population subdivision and migration is the best studied model of geographical structure in population genetics. This section treats the finite island model (Latter 1973; Maruyama 1974), in which the population is subdivided into D demes, each of size N haploid individuals, and each of which accepts a fraction m of migrants every generation. The results discussed below all hold for a diploid monoecious population if N is replaced by $2N$. The application of the island model is limited because it does not in fact contain explicit geography: migrants are equally likely to have come from any deme in the population. Therefore it cannot make a prediction of "isolation by distance" (Wright 1943), although generalized versions of the island model can (Wakeley and Aliacar 2001). The model does predict greater levels of relationship among individuals from the same deme than among individuals from different demes, and thus violates the fundamental assumption of the coalescent, that lineages are exchangeable. In the island model, rates of coalescence tend to be higher within than between demes.

Although the approximations below can be made for more general models of subdivision, the finite island model is complicated enough to illustrate the various simplifications that have been studied. Consider a sample of size two taken from the population. Larger samples can be treated using the same methods, but as with the model a limited sample is enough to illustrate the results. Generations are assumed to be nonoverlapping. At the beginning of each generation, individuals in each deme contribute a large number of "gametes" to their own deme's gamete pool and to a migrant gamete pool. Reproduction occurs within demes according to the Wright-Fisher model, except that a fraction m of gametes are sampled from the migrant pool, the other fraction coming from the deme's own gamete pool. The samples, or the ancestral lineages of the sample, can be in either of two states: (1) in the same deme or (2) in different demes. The only other possible state (3) is that the ancestral lineages of the sample have coalesced. The ancestry of the sample is a discrete-time Markov process with the following single-generation transition matrix:

$$\Pi = \begin{pmatrix} \left[\alpha + \frac{1-\alpha}{D}\right]\left(1 - \frac{1}{N}\right) & (1-\alpha)\left(1 - \frac{1}{D}\right) & \left[\alpha + \frac{1-\alpha}{D}\right]\frac{1}{N} \\ \frac{1-\alpha}{D}\left(1 - \frac{1}{N}\right) & 1 - \frac{1-\alpha}{D} & \frac{1-\alpha}{DN} \\ 0 & 0 & 1 \end{pmatrix}, \tag{2}$$

in which $\alpha = (1 - m)^2$ is the probability that neither lineage is a migrant.

The entries in Π are the probabilities of moving between states, or of staying in the same state, in a single generation looking back. For example, $(\Pi)_{13}$ is the probability of coalescence (state 3) in a single generation, given the two lineages are in the same deme now (state 1). It is equal to the probability that the lineages came from the same deme, either by staying in the same deme (with probability α) or by migrating and having the same source deme (with probability $(1 - \alpha)/D$), and that they are derived from the same parent within that deme (with probability $1/N$). State 3 is an absorbing state—once in state 3, there is zero chance of moving to states 1 or 2—and the process is followed back to the first occurrence of this, which is the most recent time the samples shared a common ancestor. The goal in analyzing this model is to obtain the t-generation transition matrix $\Pi(t) = \Pi^t$. Then $(\Pi(t))_{13}$ and $(\Pi(t))_{23}$ are, respectively, the distribution of the time to coalescence for a sample of size two from the same deme and the distribution for a sample from two different demes.

Although for the matrix Π above it is possible to obtain $\Pi(t)$ fairly easily by finding the eigenvalues and eigenvectors of the matrix, the result (not shown) is still complicated compared to the simplicity of the unstructured coalescent. Further, in the case of samples larger than size two, the matrix becomes larger and the algebra becomes intractable when the sample size is greater than about five. The complexity of many natural populations may be irreducible beyond this, and may in fact be much more complicated than the finite island model. However, there are a number of special cases of the above model which share the simplicity of the coalescent. Several of these still capture the essence of island-model subdivision, that is, greater relatedness within than between demes, while others collapse to the unstructured case. Whether these simpler versions of the model are appropriate for any particular natural population is an empirical question. Some of the results are easily obtained, while others rely on a theorem due to Möhle (1998) for Markov processes with two time scales that is detailed below in the section on low migration.

The High-Migration Limit

A somewhat trivial, introductory example is the case in which $m = 1$, that is, when individuals have no homing tendency at all. The transition matrix of Equation (2) reduces to

$$\Pi = \begin{pmatrix} \frac{1}{D}\left(1-\frac{1}{N}\right) & \left(1-\frac{1}{D}\right) & \frac{1}{DN} \\ \frac{1}{D}\left(1-\frac{1}{N}\right) & \left(1-\frac{1}{D}\right) & \frac{1}{DN} \\ 0 & 0 & 1 \end{pmatrix}. \quad (3)$$

The population is of course exactly panmictic when $m = 1$, so that all members of every deme are equally likely to have come from any deme in the population. Reproduction is population wide, and the only remnant of subdivision is that individuals reside ephemerally in demes each generation. Thus the first two rows of Π are identical; the coalescent process for a sample from the same deme is identical to the coalescent process for a sample from different demes. The probabilities in these first two rows can be obtained by imagining tossing two balls (lineages) randomly into D bins (demes) each containing N boxes (potential parents).

The matrix of Equation (3) specifies that the time to common ancestry for a pair of lineages will be geometrically distributed with mean, in generations, equal to the total population size, ND. This is identical to the result for $n = 2$ in the panmictic model. With the further assumption that ND is large, and if time is measured in units of ND generations, the distribution of the time to common ancestry for the two lineages becomes exponential as in Equation (1).

The Low-Migration Limit

The low-migration limit has been studied from a genealogical standpoint by Takahata (1991) and Notohara (2001), and by Slatkin (1981) using a forward time approach. As m gets closer and closer to zero, the probability that neither lineage migrates becomes $\alpha = (1-m)^2 \approx 1 - 2m$. The transition matrix can be written as the sum $\Pi = \mathbf{A} + m\mathbf{B}$, where

$$\mathbf{A} = \begin{pmatrix} \left(1-\frac{1}{N}\right) & 0 & \frac{1}{N} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (4)$$

and

$$\mathbf{B} = \begin{pmatrix} -2\left(1-\frac{1}{D}\right)\left(1-\frac{1}{N}\right) & 2\left(1-\frac{1}{D}\right) & -2\left(1-\frac{1}{D}\right)\frac{1}{N} \\ \frac{2}{D}\left(1-\frac{1}{N}\right) & -\frac{2}{D} & \frac{2}{DN} \\ 0 & 0 & 0 \end{pmatrix}. \quad (5)$$

If the migration rate was actually equal to zero, then $\Pi = \mathbf{A}$ and lineages in different demes would never coalesce, since $(\mathbf{A})_{22} = 1$, while lineages in the same deme would follow the usual ancestral process for the Wright-Fisher model and have a chance $1/N$ of coalescing each generation. The entries in the second row of the matrix $\mathbf{B}$ are important because they represent the chance that two separated lineages enter the same deme and thus might coalesce. Because of this, the time scale of the coalescent process will depend on m. For example, the rate $(\Pi)_{21} = m(\mathbf{B})_{21}$ at which two separated lineages enter the same deme is small if the migration probability m is small, so the time it takes for this to occur will be very long if m is close to zero.

The above is precisely the situation in which Möhle's (1998) theorem may be applied to find a continuous-time limit of a discrete-time process with events occurring on two time scales: fast in matrix $\mathbf{A}$ and slow in matrix $m\mathbf{B}$. The result is then considered an approximation for populations in which the migration rate is small. In technical terms, we define $\mathbf{A} = \lim_{m\to 0} \Pi$ and $\mathbf{B} = \lim_{m\to 0} (\Pi - \mathbf{A})/m$, and the theorem requires that the matrix $\mathbf{P} = \lim_{t\to\infty} \mathbf{A}^t$ exists. This equilibrium matrix $\mathbf{P}$ is simply the result of letting the fast process described by $\mathbf{A}$ run to its conclusion, which in this case would be guaranteed coalescence starting from state 1 and no change starting from states 2 and 3. Then, if time is measured in units of $1/m$ generations, the ancestral process is determined by the rate matrix $\mathbf{G} = \mathbf{PBP}$ and includes both the process described by $\mathbf{B}$ and the now instaneous jumps represented by the matrix $\mathbf{P}$. In particular, $\Pi(t) = \mathbf{P}e^{\mathbf{G}t}$ (Möhle 1998).

Here,

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad (6)$$

and the rate matrix simplifies to

$$\mathbf{G} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -\frac{2}{D} & \frac{2}{D} \\ 0 & 0 & 0 \end{pmatrix}, \quad (7)$$

so that, finally,

$$\Pi(t) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & e^{-\frac{2t}{D}} & 1 - e^{-\frac{2t}{D}} \\ 0 & 0 & 1 \end{pmatrix}. \quad (8)$$

Therefore, in the low-migration limit and with time measured in units of $1/m$ generations, a sample of two sequences from the same deme coalesces immediately. In truth, the time this takes will be approximately geometrically distributed with mean N generations, but this amount of time is negligible on the time scale of $1/m$ generations with $m \to 0$, so that N does not even appear in Equation (8). The distribution of time to common ancestry for a pair of sequences from different demes is exponentially distributed on this new timescale, with rate $2/D$, because there are two lineages and the chance to enter the same deme is inversely proportional to D. Note that we could rescale time again, by this factor $2/D$, and the result would be Kingman's coalescent for among-deme samples, with instantaneous coalescence of within-deme samples.

The Strong-Migration Limit

The case of strong migration was originally studied by Nagylaki (1980) in the context of the forward time diffusion of allele frequencies, and more recently by Notohara (1990) using a genealogical approach. Möhle's (1998) theorem can again be used, but now with time measured in units of N generations and letting N go to infinity for constant values of m and D. The intermediate matrices—**A**, **B**, **P**, and **G**—are not shown, only the final result:

$$\Pi(t) = \begin{pmatrix} \frac{1}{D}e^{-\frac{t}{D}} & \left(1-\frac{1}{D}\right)e^{-\frac{t}{D}} & 1-e^{-\frac{t}{D}} \\ \frac{1}{D}e^{-\frac{t}{D}} & \left(1-\frac{1}{D}\right)e^{-\frac{t}{D}} & 1-e^{-\frac{t}{D}} \\ 0 & 0 & 1 \end{pmatrix}. \quad (9)$$

The result is similar to the high-migration limit in that the genealogical process does not depend on the sampling scheme. The distribution of the time to common ancestry is exponential, as in the standard coalescent, the only difference being the measurement of time. If time in Equation (9) is rescaled again, by D, so that the units were ND generations, then the rate of coalescence would be equal to one for a sample of size two, just as in Equation (1).

The migration parameter m is no longer part of the equation. This is because, when N is large and m is not necessarily small, the lineages will have migrated so many times before they coalesce that the population will appear to be panmictic. The distribution of the two lineages among the demes reaches a statistical equilibrium so that the probability both are in the same deme is a constant $1/D$, the factor multiplying the terms in the first column of Equation (9). Note that the strong-migration limit is different than the high-migration limit, because in the strong-migration limit it is the difference in time scale between migration and coalescence, which makes the structure disappear, while in the high-migration limit there really is no structure. One can think of the strong migration limit as a reflection of the fact, discovered by Wright (1931) and illustrated in the next section, that patterns of population subdivision depend on the product Nm, and in the strong-migration limit $Nm \to \infty$.

The Structured Coalescent

This is the limit typically applied in population genetics, dating back to Wright (1931). It is appropriate when m is small and N is large, so that the effects of migration depend only on the product Nm. The structured coalescent is implicit in the work of Hey (1991), Slatkin (1987), and Strobeck (1987), with formal work by Notohara (1990) and a rigorous proof by Herbots (1994); see also Wilkinson-Herbots (1998).

Defining a new parameter M to be equal to $2Nm$, and assuming that N is large, the single-generation transition matrix becomes

$$\Pi \approx \begin{pmatrix} 1-\left[1+M\left(1-\frac{1}{D}\right)\right]\frac{1}{N} & M\left(1-\frac{1}{D}\right)\frac{1}{N} & \frac{1}{N} \\ \frac{M}{DN} & 1-\frac{M}{DN} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad (10)$$

where the approximation is that terms involving $1/N^2$ and $1/N^3$ have been dropped. Considering the limit as N goes to infinity, this Π does not include processes acting on different time scales; all changes between states occur at rates proportional to $1/N$. Thus the application of Möhle's (1998) theorem provides no simplification. There is a continuous-time approximation in which time is measured in units of N generations, which can be written $\Pi(t) = e^{\mathbf{G}t}$, where

$$\mathbf{G} = \begin{pmatrix} -1-M\left(1-\frac{1}{D}\right) & M\left(1-\frac{1}{D}\right) & 1 \\ \frac{M}{D} & -\frac{M}{D} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad (11)$$

but it is no simpler than the direct analysis of Equation (10) or even Equation (2). However, for sample sizes larger than two, the stuctured coalescent is simpler that the discrete-time model since coalescent events occur singly in the structured coalescent, whereas multiple coalescent events can occur in a single generation in the discrete-time model.

The Many-Demes Limit

The structured coalescent is a model of nontrivial population subdivision. That is, the distribution of the time to coalescence depends on the sample configuration under the structured coalescent, while in two of the three previous limits—high migration and strong migration—the genealogical process becomes the same for every kind of sample. Like the structured coalescent, the many-demes limit for the matrix in Equation (2) exhibits a nontrivial population structure, but it is also closely related to the unstructured coalescent. The many-demes limit was studied in Wakeley (1998) using a genealogical approach and in Wakeley (2003) forward in time. It is an approximation for populations with a large number of demes, and thus sits somewhere between the finite island model (Latter 1973; Maruyama 1974) and the infinite island model (Wright 1931).

The simplification again results from the application of Möhle's (1998) theorem, in the limit as $D \to \infty$ in Equation (2) and time is measured in units of D generations. The matrix $\mathbf{A}$ contains rates for coalescent events and migration events that do not bring two lineages together into the same deme, while the matrix $\mathbf{B}/D$ contains rates for migration events that do bring two lineages together into the same deme. In this case,

$$\mathbf{P} = \begin{pmatrix} 0 & 1-F & F \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad (12)$$

where

$$F = \frac{(1-m)^2}{Nm(2-m)+(1-m)^2} \quad (13)$$

is the probability that two lineages currently in the same deme coalesce before they are separted by migration. Thus F is the equivalent to one way that F_{ST} (Wright 1951) has been defined (Charlesworth 1998; Slatkin 1991). The matrix $\mathbf{G} = \mathbf{PBP}$ is readily obtained (not shown), and finally

$$\Pi(t) = \begin{pmatrix} 0 & (1-F)e^{-ct} & 1-(1-F)e^{-ct} \\ 0 & e^{-ct} & 1-e^{-ct} \\ 0 & 0 & 1 \end{pmatrix}, \quad (14)$$

in which

$$c = \frac{m(2-m)}{Nm(2-m) + (1-m)^2} = \frac{1-F}{N}, \qquad (15)$$

describes the ancestral process for a sample of two lineages when time is measured in units of D generations and D is large. Thus, in the many-demes limit, the time to common ancestry for a sample of two sequences from two different demes is exponentially distributed with rate c on this time scale. If time is measured in units of $ND/(1 - F)$ generations, then the rate becomes one just as in Kingman's coalescent. A sample of two sequences from the same population has an initial chance F of coalescing (at $t = 0$), and with chance $1 - F$ it has an exponentially distributed coalescence time identical to that of a single-deme sample.

Discussion

All of the limits discussed above can be extended to samples larger than two. In the high-migration limit and the strong-migration limit, the result is always complete collapse to the unstructured coalescent. The structured coalescent retains its complexity, and it becomes necessary to model the locations of all the lineages back in time. The low-migration limit and the many-demes limit become more general versions of the two-phase processes described above. In both cases, the history of a sample of sequences taken singly from different demes follows an unstructured coalescent model, but with an effective size that is different than the census size ND of the population. In the low-migration limit and the many-demes limit, this effective size depends inversely on the migration rate because migration is the process that brings lineages into the same deme so they can coalesce. Sample configurations in which the sample size is greater than the number of sampled demes have two parts to their history. First, there is an initial burst of coalescent events for within-deme samples, and possibly some migration events, before the remaining lineages, which are all now in separate demes, enter the unstructured coalescent process. In the many-demes model, these have been respectively called the "scattering" phase and the "collecting" phase in consideration of the role of migration during each (Wakeley 1999). In the low-migration limit, all samples from a single deme will coalesce to a single lineage in the scattering phase.

The limits above can also be extended to more general population models, including population structures in which demes differ in size and migration rate, and in which migration is not necessarily equally probable for every pair of demes. In the face of this, the complexity of the structured coalescent increases quickly, while the other limits remain functions of a much smaller number of parameters due to their connection, via an effective population size, with the unstructured coalescent. For example, histories under the low-migration limit depend only on this effective size since all within-deme samples coalesce during the scattering phase. In the many-demes model, the history of the sample depends directly on the parameters for the sampled demes, while the only effect of the many unsampled demes is through the effective size. This is in contrast to the case of the structured coalescent, in which the effects of unsampled demes are not captured in an effective population size, and in which it is typically assumed in applications that the sampled demes constitute the entire population; but see Beerli (2004).

Given the current ease of sequencing DNA and the continued improvements in biotechnology, large multilocus datasets will be the norm in population genetics studies in the coming years for nonmodel as well as model organisms. Even now, computational methods of inference and analytical work on the necessary models do not meet the needs of researchers, so there should be continued effort in both these subfields of population genetics. The results summarized here show that complex demographic scenarios can, in some cases, be described using relatively simple models. Which of these models, if any, is appropriate for a particular population is an empirical question, and should be considered separately from the ease with which these models can be applied. Populations with small numbers of demes, small migration rates, and large deme sizes will require the complexity of the structured coalescent. In the simpler cases, the effect of structure either (1) disappears entirely, as in the high-migration and strong-migration limits, or (2) reduces to separable effects on the time scale of coalescence and on levels of within-deme versus between-deme relatedness, as in the low-migration limit and the many-demes limit. Other behaviors are possible in other kinds of populations, and the methods reviewed here should aid in the derivation of results in a variety of situations.

Acknowledgment

This paper was originally presented at the American Genetics Association 2003 Annual Meeting and Centennial Celebration at the University of Connecticut, Storrs, July 18–30, 2003.

References

Aldous DJ, 1985. Exchangeability and related topics. pp. 1–198. In: *École d'Été de Probabilités de Saint-Flour XII–1983*, vol. 1117 of *Lecture Notes in Mathematics*. (Dold A and Eckmann B, eds). Berlin: Springer-Verlag.

Beerli P, 2004. Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Mol Ecol* 13:827–836.

Cannings C, 1974. The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. *Adv Appl Prob* 6:260–290.

Charlesworth B, 1998. Measures of divergence between populations and the effect of forces that reduce variability. *Mol Biol Evol* 15:538–543.

Donnelly P and Tavaré S, 1995. Coalescents and genealogical structure under neutrality. *Annu Rev Genet* 29:401–421.

Ewens WJ, 1972. The sampling theory of selectively neutral alleles. *Theor Popul Biol* 3:87–112.

Ewens WJ, 1990. Population genetics theory – the past and the future. In: *Mathematical and statistical developments of evolutionary theory* (Lessard S, ed). Amsterdam: Kluwer Academic; 177–227.

Fisher RA, 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Trans Soc Edinb* 52:399–433.

Fisher RA, 1930. *The genetical theory of natural selection*. Oxford: Clarendon.

Griffiths RC, 1979. Exact sampling distributions from the infinite neurtal alleles model. *Adv Appl Prob* 11:326–354.

Griffiths RC, 1980. Lines of descent in the diffusion aproximation of neutral Wright-Fisher models. *Theor Popul Biol* 17:37–50.

Haldane JBS, 1932. *The causes of natural selection*. London: Longmans Green & Co.

Harpending H, Batzer MA, Gurven M, Jorde LB, Rogers AR, and Sherry ST, 1998. Genetic traces of ancient demography. *Proc Natl Acad Sci USA* 95:1961–1967.

Harris H, 1966. Enzyme polymorphism in man. *Proc R Soc Lond B* 164:298–310.

Hawks J, Hunley K, Lee S-H, and Wolpoff M, 2000. Population bottlenecks and Pleistocene human evolution. *Mol Biol Evol* 17:2–22.

Herbots HM, 1994. *Stochastic models in population genetics: genealogical and genetic differentiation in structured populations* (PhD dissertation). London: University of London.

Hey J, 1991. A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. *Theor Popul Biol* 39:30–48.

Hudson RR, 1983a. Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol* 23:183–201.

Hudson RR, 1983b. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* 37:203–217.

Hudson RR, 1990. Gene genealogies and the coalescent process. In: *Oxford surveys in evolutionary biology*, vol. 7. (Futuyma DJ and Antonovics J, eds). Oxford: Oxford University Press; 1–44.

International SNP Map Working Group, 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–933.

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

Kaplan NL and Hudson RR, 1985. The use of sample genealogies for studying a selectively neutral *m*-loci model with recombination. *Theor Popul Biol* 28:382–396.

Kaplan NL, Hudson RR, and Iizuka M, 1991. Coalescent processes in models with selection, recombination and geographic subdivision. *Genet Res Camb* 57:83–91.

Kaplan NL, Hudson RR, and Langley CH, 1989. The "hitchhiking effect" revisited. *Genetics* 123:887–899.

Kimura M, 1955a. Solution of a process of random genetic drift with a continuous model. *Proc Natl Acad Sci USA* 41:144–150.

Kimura M, 1955b. Stochastic processes and the distribution of gene frequencies under natural selection. *Cold Spring Harbor Sympo Quant Biol* 20:33–53.

Kimura M, 1969. The number of heterozygous nucleotide sites maintained in a finite population due to the steady flux of mutations. *Genetics* 61:893–903.

Kimura M and Crow JF, 1964. The number of alleles that can be maintained in a finite population. *Genetics* 49:725–738.

Kingman JFC, 1982a. The coalescent. *Stochastic Process Appl* 13:235–248.

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

Kingman JFC, 1982c. On the genealogy of large populations. *J Appl Prob* 19A:27–43.

Kingman JFC, 2000. Origins of the coalescent: 1974–1982. *Genetics* 156:1461–1463.

Kreitman M, 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* 304:412–417.

Krone SM and Neuhauser C, 1997. Ancestral processes with selection. *Theor Popul Biol* 51:210–237.

Latter BDH, 1973. The island model of population differentiation: a general solution. *Genetics* 73:147–157.

Lewontin RC, 1974. *The genetic basis of evolutionary change*. New York: Columbia University Press.

Lewontin RC and Hubby JL, 1966. A molecular aproach to the study of genic diversity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* 54:595–609.

Malécot G, 1946. La consaguinité dans une population limitée. *C R Acad Sci Paris* 222:841–843.

Malécot G, 1948. *Les Mathématiques de l'Hérédité*. Paris: Masson. Extended translation in *The Mathematics of Heredity*. San Francisco: WH Freeman, 1969.

Maruyama T, 1974. A simple proof that certain quantities are independent of the geographical structure of population. *Theor Popul Biol* 5:148–154.

Möhle M, 1998. A convergence theorem for Markov chains arising in population genetics and the coalescent with partial selfing. *Adv Appl Prob* 30:493–512.

Nagylaki T, 1980. The strong-migration limit in geographically structured populations. *J Math Biol* 9:101–114.

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

Nordborg M, 2001. Coalescent theory. In: *Handbook of statistical genetics* (Balding DJ, Bishop MJ, and Cannings C, eds). Chichester, England: John Wiley & Sons.

Notohara M, 1990. The coalescent and the genealogical process in geographically structured population. *J Math Biol* 29:59–75.

Notohara M, 2001. The structured coalescent process with weak migration. *J Appl Prob* 38:1–17.

Pluzhnikov A, Rienzo AD, and Hudson RR, 2002. Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics* 161:1209–1218.

Provine WB, 1971. *The origins of theoretical population genetics*. Chicago: University of Chicago Press.

Przeworski M, Hudson RR, and DiRienzo A, 2000. Adjusting the focus on human variation. *Trends Genet* 16:296–302.

Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Huggins JM, Richter DJ, Lander ES, and Altshuler D, 2002. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet* 32:135–142.

Slatkin M, 1981. Fixation probabilities and fixation times in a subdivided population. *Evolution* 35:477–488.

Slatkin M, 1987. The average number of sites separating DNA sequences drawn from a subdivided population. *Theor Popul Biol* 32:42–49.

Slatkin M, 1991. Inbreeding coefficients and coalescence times. *Genet Res Camb* 58:167–175.

Slatkin M and Hudson RR, 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129:555–562.

Stephens M, 2001. Inferences under the coalescent. In: *Handbook of statistical genetics* (Balding DJ, Bishop MJ, and Cannings C, eds). Chichester, England: John Wiley & Sons.

Strobeck C, 1987. Average number of nucleotide differences in an sample from a single subpopulation: a test for population subdivision. *Genetics* 117:149–153.

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

Takahata N, 1991. Genealogy of neutral genes and spreading of selected mutations in a geographically structured population. *Genetics* 129: 585–595.

Takahata N, 1995. A genetic perspective on the origin and history of humans. *Annu Rev Ecol Syst* 26:343–372.

Tavaré S, 2004. Ancestral inference in population genetics. In: *École d'Éte´de Probabilités de Saint-Flour XXXI – 2001*, *Lecture Notes in Mathematics*. edited by Cantoni O, Tavaré S, and Zeitouni O, eds). Berlin: Springer-Verlag.

Wakeley J, 1998. Segregating sites in Wright's island model. *Theor Popul Biol* 53:166–175.

Wakeley J, 1999. Non-equilibrium migration in human history. *Genetics* 153:1863–1871.

Wakeley J, 2003. Polymorphism and divergence for island model species. *Genetics* 163:411–420.

Wakeley J and Aliacar N, 2001. Gene genealogies in a metapopulation. *Genetics* 159:893–905 [Corrigendum (Figure 2): *Genetics* 160:1263–1264].

Wakeley J and Lessard S, 2003. Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. *Genetics* 164:1043–1053.

Watterson GA, 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* 7:256–276.

Watterson GA, 1976a. Reversability and the age of an allele. I. Moran's infinitely many neutral alleles model. *Theor Popul Biol* 10:239–253.

Watterson GA, 1976b. The stationary distribution of the infinitely many neutral alleles diffusion model. *J Appl Prob* 13:639–651.

Watterson GA, 1977. Heterosis or neutrality? *Genetics* 85:789–814.

Wilkinson-Herbots HM, 1998. Genealogy and subpopulation differentiation under various models of population structure. *J Math Biol* 37:535–585.

Wright S, 1931. Evolution in Mendelian populations. *Genetics* 16:97–159.

Wright S, 1943. Isolation by distance. *Genetics* 28:114–138.

Wright S, 1951. The genetical structure of populations. *Ann Eugenics* 15:323–354.

Corresponding Editor: Kent E. Holsinger