Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Developments in coalescent theory from single loci to chromosomes

John Wakeley

Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, 02138, USA

ARTICLE INFO

Article history:
Received 19 June 2019
Available online xxxx

Keywords:
Coalescent theory
Gene genealogies
Population genetics
Mutation models
Recombination

Contents

Coalescent theory takes a retrospective approach to population genetics in order to model genetic ancestry. The overall goal is to understand the forces which produce and maintain genetic variation. The more specific goals are to make predictions about patterns of genetic variation and to use these predictions with data to make inferences about mutation, recombination, population sizes through time, population subdivision and natural selection. In coalescent theory, genetic ancestries of samples are modeled with as little reference as possible to the remainder of the population. The amount of information needed depends on the structure and dynamics of the population, including whether there is selection. Focusing on samples is useful because all observed genetic polymorphisms result from mutations which happened in the ancestry of the sample. The genetic ancestry of a sample at single locus is called a gene genealogy or coalescent tree. Because these ancestries may trace back tens of thousands to millions of years, they cannot be observed directly. They appear as latent variables or nuisance parameters in coalescent theory. Importantly, the gene genealogies at different loci along a chromosome may differ when there is recombination.

The method of Li and Durbin (2011), in which variation in amounts of polymorphism across the genome are used to infer changes in population size over time, provides an excellent example of how coalescent theory may be applied. One clear prediction of the theory is that loci with longer gene genealogies, which extend farther back in time and thus have had more opportunity for mutation, will tend to display more genetic variation than loci with shorter gene genealogies. Another prediction is that the lengths of gene genealogies depend inversely on population size, because the chance of finding common ancestors is greater in smaller populations. Finally, recombination causes different loci in the same genome to have different gene genealogies because it makes them follow different paths through the organismal genealogy of the population. Li and Durbin (2011) used an approximation to the coalescent model with recombination, discussed in Section 6, to design a hidden Markov model of gene genealogies along the two chromosomes of a single diploid genome. From the inferred distribution of times to common ancestry across the genome, they were able to show how the ancestral population size of humans has changed over time.

E-mail address: wakeley@fas.harvard.edu.

https://doi.org/10.1016/j.tpb.2020.02.002

Please cite this article as: J. Wakeley, Developments in coalescent theory from single loci to chromosomes. Theoretical Population Biology (2020), https://doi.org/10.1016/j.tpb.2020.02.002.

In the fifty years since the founding of *TPB*, the theory of gene genealogies went from a bunch of disconnected, implicit ideas to a rigorous framework for population-genetic inference. This Commentary addresses two aspects of coalescent theory in the journal. The first is the series of fundamental contributions to the development of coalescent theory in the 1970s and 1980s. During that time, population genetics got its first major infusion of data. Theory grew to explain and to draw inferences from newly observed patterns of variation. Predictions from different models were slowly recognized as having a common structure, due to the unobserved gene genealogies behind the data. Soon it became the norm to make predictions and draw inferences by averaging over these latent variables. As DNA sequencing became easier and the prevalence of recombination was recognized, single-locus versions of the theory gave way to problems of linkage and recombination. Today, the data are genome sequences from very large numbers of individuals, and the field is in the midst of new expansion. The second aim of this Commentary is to trace the development of coalescent models of recombination, from the key initial descriptions, calculations and simulations to the growing array of current approximate methods which scale up to make inferences from many whole genomes.

The standard neutral coalescent process, including the possibility of recombination, is the main focus of this Commentary. It is an idealized model, which holds for large well-mixed populations of constant size through time and without selection. It is the statistical prior or null model against which inferences about selection, population structure, variable population size, etc., are made. Much of the history of thought about this model has taken place in *TPB*. Extensions of it – for example to account for natural selection, population structure or high variance of offspring numbers – will be mentioned briefly. For more detail see the concise reviews by Hudson (1990), Donnelly and Tavaré (1995) and Nordborg (2001), or the book-length treatments by Tavaré (2004), Hein et al. (2005), Durrett (2008), Wakeley (2008) and Berestycki (2009).

1. Fundamentals of the coalescent process

The gene genealogy of a sample is the set of ancestral genetic relationships among n gene copies or alleles at a single locus. A locus could be one nucleotide site in a genome or a stretch of sites within which there is no recombination. Hudson (1983a,b) and Tajima (1983) introduced the theory of gene genealogies in a population-genetic context whereas Kingman (1982a,b,c) established the backward-time process of coalescence in a rigorous mathematical setting. In Kingman's terminology, the 'n-coalescent' is a continuous-time Markov process on the set of equivalence relations of the integers 1 through n, which begins with each sample in its own class and ends with all samples in one class. In the limit as the population size N tends to infinity, the process involves exactly $n-1$ binary mergers of classes, or ancestral genetic lines. The outcome may be depicted as a binary tree with external nodes labeled 1 through n and internal nodes representing coalescent events ordered in time. Intuitively, gene genealogies are constructed by letting each pair of ancestral lines merge at rate 1, until the most recent common ancestor of the entire sample is reached.

Time must be rescaled in order to obtain the standard neutral coalescent model in the limit $N \to \infty$. Specifically, time is measured in units of N_e generations, where N_e is the coalescent effective population size, which is equal to the expected number of generations back to the common ancestor for a sample of size two (Möhle, 2001; Ewens, 2004; Sjödin et al., 2005). Typically N_e is proportional to the actual population size (Sjödin et al., 2005) but other possibilities exist (Wakeley and Sargsyan, 2009). In the well studied Wright–Fisher model of reproduction (Fisher, 1930; Wright, 1931), $N_e = 2N$ for monoecious diploids and $N_e = N$ for haploids. The standard neutral coalescent model is robust to phenomena which occur on time scales much shorter than N_e generations (Möhle, 1998). Thus it is expected to hold for diverse species with different mating structures and distributions of family sizes, as well as mild forms of population subdivision, as long as N_e is large. However, fundamental modifications are required to accommodate selection (Kaplan et al., 1988) or population subdivision with low rates of migration (Takahata, 1988).

The backward-time models of coalescent theory are mathematically dual to the forward-time diffusion models of population genetics (Ewens, 1990, 2004; Möhle, 2001). Both make the same assumptions and either can be used to predict patterns of genetic variation or to make inferences from data. Diffusion models make predictions about the entire population and more easily accommodate selection. Coalescent models focus on sample statistics and provide a natural framework for inferences about ancestry.

When neutral mutations occur with probability u per generation at the locus, the model is characterized by a single mutation parameter, which is defined as $\theta = 2N_e u$. What happens when a mutation occurs will depend on the kind of data being considered (e.g., see Section 2, 4, and 5) but in any case the standard neutral coalescent process with mutation may be described in terms of two overall rates. When there are i lineages ancestral to the sample, the total rates of coalescence and mutation are $i(i-1)/2$ and $i\theta/2$. Therefore, because the waiting time to a coalescent event is $2/(i(i-1))$ on average, gene genealogies have shorter times to common ancestors in the recent past when there are more ancestral lineages, and longer times in the more distant past when there are fewer. Also, the expected number of mutations during the time when there are i lineages ancestral to the sample is equal to $\theta/(i-1)$, which means that polymorphisms will tend to trace back to mutations in the more distant past, i.e. when i is smaller, and that there will be a diminishing return on discovering additional polymorphisms with increasing sample size.

The notions underlying coalescent theory had been developing for several decades, especially in studies of identity by descent for samples of size two (Wright, 1922; Malécot, 1941, 1946). But the full importance of gene genealogies became clear after patterns of variation in larger samples were revealed using the techniques of molecular biology, beginning in the 1960s. The first polymorphism data came from protein gel electrophoresis (Harris, 1966; Lewontin and Hubby, 1966). Restriction fragment length polymorphisms were reported somewhat later (Shah and Langley, 1979; Brown, 1980), followed relatively quickly by DNA sequences (Kreitman, 1983). Coalescent theory emerged in the study of mutation models tailored to these new data: from the infinite-alleles model (Malécot, 1946; Kimura and Crow, 1964) and the step-wise mutation model (Ohta and Kimura, 1973) for allelic data, to the infinite-sites models with independent sites (Kimura, 1969, 1971; Ewens, 1974) or perfectly linked sites (Watterson, 1975) in anticipation of DNA sequence data. The same year the first population-genetic sample of DNA sequences appeared, Hudson (1983b) described the coalescent process with arbitrary levels of recombination and infinite-sites mutation.

2. Infinite-alleles model: Equilibrium properties

Ewens (1972) presented the first sampling formula for genetic variation, describing the joint distribution of the number of alleles and their frequencies in a sample of size n. Karlin and McGregor (1972) proved Ewens' formula using a recursive equation which is exact for the coalescent process (Kingman, 1982b) and which sums over the possible mutation events and coalescent events affecting the sample. Ewens (1972) marks the start of modern-day population genetics, in which intricate patterns of polymorphism are analyzed using sophisticated statistical machinery to make inferences about effective population sizes, mutation and recombination rates, population structure and selection. The Ewens sampling formula went a long way toward explaining patterns of electrophoretically detectable variation, despite the fact that it was based on assumptions of selective neutrality and infinite-alleles mutation. It predicted what Lewontin and Hubby (1966) and many others subsequently had observed, that samples tend to contain many low-frequency alleles with one or perhaps two alleles segregating in higher frequency. As a null model, it provided a means of estimating the population-scaled mutation rate, θ, and a framework for testing the neutral mutation hypothesis against alternatives which include selection (Ewens, 1972; Watterson, 1977, 1978).

It took some time to understand the temporal structure behind Ewens' formula. Working at first outside population genetics, Kingman (1975) introduced the Poisson–Dirichlet distribution, which also applies to allele frequencies in a population under infinite-alleles mutation when the frequencies are ordered largest to smallest (Watterson, 1974, 1976b). Watterson and Guess (1977) connected age order to frequency by showing that the probability the most common allele is the oldest allele in the population is equal to its frequency. Watterson (1976b) obtained an explicit formula for the Poisson–Dirichlet distribution and proved that samples from it obey the Ewens sampling formula. Kingman (1977b) proved that allele frequencies in the total population must be Poisson–Dirichlet distributed when Ewens' formula holds for samples. Arratia et al. (2003) describe the generality and subsequent wide application of these two related distributions. By the late 1970s, a thorough description of equilibrium patterns of infinite-alleles variation was available. It made some connections to temporal structure, which were further elaborated in the theory of lines of descent (see Section 3) then later drawn out in detail when Kingman (1982b) showed that the Ewens sampling formula follows immediately from the structure of gene genealogies plus infinite-alleles mutation.

Ewens (1972) and Karlin and McGregor (1972) had focused on the probability that the next sample taken is of a novel allelic type. For the nth sample, this probability is $\theta/(\theta+n-1)$. The alternative, that the nth sample is of a type which has already been observed, has probability $(n-1)/(\theta+n-1)$. In the latter case, it is equally likely to be any of the types already observed. Under the new temporal interpretation involving coalescence backward in time, these two probabilities appear as ratios of rates. Thus,

$$\frac{n\theta/2}{n\theta/2+n(n-1)/2}=\frac{\theta}{\theta+n-1}$$

is the probability that the first event back in the ancestry of the sample is a mutation event. Similarly, $(n-1)/(\theta+n-1)$ is the probability that the first event back in the ancestry is a coalescent event. Keeping track of all alleles and their sample frequencies, the full Ewens sampling formula can be derived either by modeling successive samples or by following all samples' ancestries backward in time. For example, the probability that no mutations happen on the entire gene genealogy is given by

$$\frac{(n-1)!}{(\theta+1)\cdots(\theta+n-1)}$$

which is identical to equation (19) in Ewens (1972) for the probability that all n samples are of the same allelic type, except that Ewens' equation (19) is undefined when $\theta=0$.

Based on these probabilities, $\theta/(\theta+n-1)$ and $(n-1)/(\theta+n-1)$, Hoppe (1984, 1987) described a Pólya-like urn model for successive sampling, which produces the Ewens sampling formula and is equivalent to a forward-time description of the coalescent model with mutation as a Markov jump process. Similar generative models for sampling probabilities hold, alongside backward-time coalescent descriptions, for diffusion approximations to neutral population models with a variety of types of mutation (Donnelly, 1986; Ethier and Griffiths, 1987; Donnelly and Kurtz, 1996a). In fact, the whole of diffusion theory in population genetics may be recast using particle representations which encompass the genealogical relationships in the population (Donnelly and Kurtz, 1996b, 1999; Etheridge and Kurtz, 2019). The look-down construction of Donnelly and Kurtz (1996b), in which particles (i.e. gene copies or haploid individuals) occupy levels and a newly produced particle at level n is either mutated or "looks down" to find its parent among the $n-1$ particles below it, places the Hoppe urn model and its generalizations in this more general context.

3. Infinite-alleles model: Shifting lines of descent

Just prior to the introduction of coalescent theory, a closely related forward-time theory of lines of descent was developed by Griffiths (1980). A line of descent is an inverted tree encompassing all the descendants of a root lineage (itself defined by a unique mutation) which have not subsequently mutated (Griffiths, 1980). Under the infinite-alleles model, each mutation removes one lineage from some line of descent and starts a new line of descent. Griffiths (1979) had used this notion to obtain a time-dependent version of the Ewens sampling formula. Watterson (1984) made further developments and discussed points of contact between lines of descent and coalescence. The forward-time theory of lines of descent is quite close to the backward-time analysis of ancestral lineages with mutation. For example, equation (11) in Griffiths (1980) gives the expected length of time during which there are l lines of descent

$$2/[l(\theta+l-1)]$$

which is identical to the expectation of the time to the first event in the ancestry of l lineages when the total rates of coalescence and mutation are $l(l-1)/2$ and $l\theta/2$. It may be helpful to recall Theorem 2 of Watterson (1976a) concerning the equilibrium behavior of the model of Moran (1958) with infinite-alleles mutation, which states that the ages of alleles have the same joint distribution as their extinction times. This and a number of related results follow from the reversibility of the Moran model and the corresponding diffusion (Watterson and Guess, 1977; Kelly, 1979; Donnelly, 1986).

The comprehensive synthetic work of Tavaré (1984) placed the theories of coalescence, lines of descent and ages of alleles within a single framework. A key part of this was to show that properties of the ancestral process of gene genealogies are obtained from allelic models in the limit as θ tends to zero. This highlighted the earlier work of Felsenstein (1971) which established a recursive equation for sampling probabilities of numbers of alleles at two different time points in the absence of mutation. Felsenstein employed a matrix, G, the (i,j)th entry of which is the probability that i lineages have j ancestors in the previous generation. Longer times are treated by taking powers of G. Felsenstein (1971) considered the probability that i alleles present in the population now will all still be present at some future time. He studied the leading-order, or asymptotic, rate of decay of this probability. Kimura (1955) had shown previously using diffusion theory that this rate is equal to $i(i-1)/2$ on the diffusion time scale. Felsenstein (1971) showed that a genealogical approach based on G gives the same answer, and that $i(i-1)/2$ is also the rate of decay of the probability that i alleles are present in a sample of size i.

Please cite this article as: J. Wakeley, Developments in coalescent theory from single loci to chromosomes. Theoretical Population Biology (2020), https://doi.org/10.1016/j.tpb.2020.02.002.

For the haploid Wright–Fisher model, equations (14) and (15) in Felsenstein (1971) give the probability there are still i distinct lineages left after one unit of time on the diffusion (or coalescent) time scale:

$$\lim_{N\to\infty}(G_{ii})^N = e^{-i(i-1)/2}.$$

In addition, Felsenstein's equation (28)

$$\lim_{N\to\infty}[(1-G_{ii})/(1-G_{22})] = i(i-1)/2$$

established a link between the Wright–Fisher model and the Moran model, effectively rescaling time by the expected number of generations back to the common ancestor for a sample of size two, here $1/(1-G_{22})$. Felsenstein (1971) went on to speculate about the breadth of application of this result, offering a toy model as an exception which may be seen as anticipating later work on multiple-mergers coalescent processes (Pitman, 1999; Sagitov, 1999; Schweinsberg, 2000; Möhle and Sagitov, 2001; Birkner et al., 2005).

Following Kimura (1955), Felsenstein (1971) considered these results in relation to the rate of loss of i alleles at some distant future time. But it is remarkable how close this came to the backward-time coalescent process without mutation, in which $i(i-1)/2$ is the total rate of coalescence when there are i ancestral lineages. Felsenstein (1971) did not consider what would now be called the branching structure of the gene genealogy. The rate $i(i-1)/2$ specifies a coarsened version of Kingman's 'n-coalescent', namely the 'death process' (Kingman, 1982a; Tavaré, 1984) which counts only the number of ancestral lineages backward in time. Note that a variety of intermediate resolutions are possible between this and the full-blown 'n-coalescent' (Sainudiin et al., 2015) including the 'evolutionary relationships' of Tajima (1983) which are gene genealogies with coalescent events ordered in time but with external nodes unlabeled.

4. Step-wise mutation and common ancestry

Ohta and Kimura (1973) proposed the charge-state model or step-wise mutation model for electrophoretic data as an alternative to the infinite-alleles model. In the simplest version of this model, each mutation is equally likely to add $+1$ or -1 to the state of an allele. Thus, the mutation process along a single lineage is equivalent to a random walk on the integers. The difference in positions of two alleles carries some information about their relatedness but can also be zero even if multiple mutations have occurred. It was soon discovered that mutations affecting electrophoretic mobility are more varied than unit changes in charge (Ramshaw et al., 1979; Fuerst and Ferrell, 1980). Later, step-wise mutation models were resurrected for use with repeat loci such as microsatellites (Valdes et al., 1993; Slatkin, 1995; Goldstein et al., 1995).

It is well known that a group of independent random walkers will spread out indefinitely over time. In contrast, a population of reproducing organisms undergoing stepwise mutation will wander allelic space together in a clump due to their relatedness. There is nothing comparable in elegance to the Ewens sampling formula in this case, except for a sample of size two. Ohta and Kimura (1973) found

$$1/\sqrt{2\theta+1}$$

for the probability that two samples are of the same allelic type. As before $\theta = 2N_e u$, but now when a mutation occurs it changes the state of the allele by $+1$ or -1 with equal probability. When θ is small, $1/\sqrt{2\theta+1}$ is close to the corresponding result for infinite-alleles mutation, $1/(\theta+1)$, otherwise it is smaller due to the possibility of convergence of allelic types under step-wise mutation. Moran (1975, 1976) showed that the variance of allelic values among individuals in the population is equal to θ, and obtained an expression for the distribution of pairwise differences for $n = 2$. Kingman (1976, 1977a) established a recurrence relation for the full sampling distribution for arbitrary n, and showed that the Ewens sampling formula can be obtained as a limit of a multi-dimensional step-wise model. Kingman (2000) cites these results, together with comments from Ewens about their being due to relatedness, as early hints that there was an interesting underlying mathematical object to study, namely the gene genealogy of the sample.

5. Infinite-sites mutation and gene genealogies

Today DNA sequences comprise the bulk of data in population genetics. The best sort of model for these data would be four-state mutation models which account for differences in mutation rates among the four nucleotides and for the possibility of multiple mutations at single sites (Jenkins et al., 2014; Burden and Tang, 2017). But polymorphisms are fairly rare in most species. The average number of pairwise nucleotide differences falls between 5 in 100 and 5 in 10000 across a broad range of taxa (Leffler et al., 2012). The infinite-sites model assumes that multiple mutations at single sites do not happen; every mutation occurs at a previously un-mutated site (Kimura, 1969, 1971). This will be a poor approximation for the long divergences between species but may be a good starting point for polymorphism within species. Humans are among the least diverse species (Leffler et al., 2012). A coalescent model with infinite-sites mutation and $\theta = 1$ might correspond to a locus of roughly 2000 nucleotide sites in the human genome (Tian et al., 2019).

Ewens (1974) studied infinite-sites mutation under the assumption that sites assort independently, or recombine freely. Using the allele-frequency spectrum for the corresponding infinite-sites diffusion model, he obtained the expression for the expected number of segregating sites in a sample of size n

$$E[S] = \theta\sum_{i=1}^{n-1}\frac{1}{i}$$

which Watterson (1975) found too assuming no recombination, and which illustrates the diminishing return of additional sampling mentioned in Section 1. Ewens' approach makes it clear that the individual terms of $E[S]$, namely θ/i for $i \in (1,\dots,n-1)$ may be interpreted as expected numbers of sites at which the derived allele, or mutant base, is found in i copies in the sample (the 'site-frequency spectrum'). Fu (1995) did the seminal coalescent work on the expectations, variances and covariances of site-frequency counts. Extensions have been made to changing population size (Griffiths and Tavaré, 1998; Polanski and Kimmel, 2003; Polanski et al., 2003; Evans et al., 2007), to two-site patterns (Sargsyan, 2015; Ferretti et al., 2018) and recently to the likelihood of the full site-frequency spectrum of a sample at a locus without recombination (Sainudiin and Véber, 2018). Ewens (1974) noted that S is Poisson-distributed under free recombination; that $E[S]$ holds regardless of the level of recombination; and that independence of sites gives a lower bound on $\mathrm{Var}[S]$ and therefore an upper bound on the efficiency of estimates of θ from the number of segregating sites.

Watterson (1975) studied the distribution of S at a locus without recombination. For reference, S used here corresponds to K_n in Watterson (1975). In addition to the expression for $E[S]$, he obtained

$$\mathrm{Var}[S] = \theta\sum_{i=1}^{n-1}\frac{1}{i} + \theta^2\sum_{i=1}^{n-1}\frac{1}{i^2}$$

which is the largest possible value of this variance because sites within a locus are most highly correlated when there is no recombination. In fact, Watterson (1975) showed how the full distribution of S can be computed, and not just for this case of a relatively small sample from a very large population but also for large samples, up to the total population, and for populations of any size. He took the genealogical-matrix approach of Felsenstein (1971) to a new level, placing it within the established framework of occupancy problems (Johnson and Kotz, 1969) and noting that an explicit formula for G_{ij} under the Wright–Fisher model as well as large-sample asymptotic results supporting further analysis were available. In the case of relatively small sample of size n, Watterson (1975) recognized that the ancestry would involve $n-1$ intervals, times during which the sample would have $i \in (n, n-1, \ldots, 2)$ ancestral lineages, and further that during each of these intervals some random number of mutations would occur. He found the now familiar geometric distribution of the number of mutations during the time when there were i ancestral lineages

$$\frac{i-1}{\theta+i-1}\left(\frac{\theta}{\theta+i-1}\right)^{k} \qquad k = 0, 1, 2, \ldots$$

which depends on the two ratios of rates of coalescence and mutation discussed in Section 2. The difference is that here, under the infinite-sites model, every mutation will be observed as a segregating site.

Thus, Watterson (1975) was the first to present gene genealogies and their backward-time construction, through a series of $n-1$ independent intervals and with the familiar random scattering of neutral mutations on the branches the gene genealogy, in a way that unambiguously captures our modern notion of coalescent theory. Key aspects of the theory which are missing in Watterson (1975) compared to Kingman (1982a,b,c) are the description of the detailed relationships among n labeled samples, that is the state space of gene genealogies, and the proof of convergence to the coalescent process. Following the timely and revealing article by Dung et al. (2019), it is important to recognize the contributions of Margaret Wu to Watterson (1975) which has had such an enormous impact on population genetics.

Watterson (1975) obtained the distribution of S, but there is no closed-form expression for the probability of a full data set of sequences, even under the relatively simple infinite-sites model without intra-locus recombination. Sequence data include not only the number and sample frequencies of mutations but also how these mutations are distributed among the sequences. If data from an infinite-sites locus without recombination were summarized as the number of distinct sequences, or haplotypes, and their sample frequencies, then the Ewens sampling formula would apply because Watterson's model is an infinite-alleles model. But it is of interest – both theoretically and for use in likelihood-based methods of inference – to be able to compute the probability of a full data set, including the specific mutation patterns which define the haplotypes.

For the case of no intra-locus recombination, this problem was solved in the 1990s using two different methods of integrating over the underlying, unknown gene genealogy. Details of these methods can be found in Tavaré (2004) and Wakeley (2008). Briefly, the first method averaged over series of mutation events and coalescent events in the ancestry of the sample. Strobeck (1983) proposed an early version of this for $n = 3$. Later, using the general recursive equations of Ethier and Griffiths (1987) for sampling probabilities in the infinite-sites model, Griffiths and Tavaré (1994a,b) introduced a Monte Carlo method of averaging over series of events which might have produced the data. This is a type of importance sampling (Felsenstein et al., 1999). Stephens and Donnelly (2000) made use of the look-down construction (Donnelly and Kurtz, 1996b) described in Section 2 to suggest a better proposal distribution for importance sampling. More recently, taking the forward-time generative approach, Wu (2010) developed a dynamic-programming method to solve the Ethier–Griffiths–Tavaré recursions exactly. The second method of computing likelihoods averaged over gene genealogies explicitly. In particular, Kuhner et al. (1995) introduced a Markov chain Monte Carlo (MCMC) algorithm to traverse the space of gene genealogies, including both tree shapes and coalescence times. Dealing with coalescent trees has the advantage that finite-sites mutation models, such as four-state models for DNA, can be accommodated easily. A variety of MCMC methods of inference have been developed for a wide range of extensions of the standard neutral coalescent model, and are included in the software packages LAMARC 2.0 (Kuhner, 2006) and BEAST (Drummond et al., 2012).

6. From loci to chromosomes: Recombination

At the outset of coalescent theory, Hudson (1983b) provided a full description of gene genealogies in the presence of recombination. The single-locus theories described above are insufficient because the rate of recombination is similar to the rate of mutation across a wide array of taxa; see Table 4.1 in Lynch (2007). Even though recombination varies along the genome and is concentrated in hotspots (Lichten and Goldman, 1995; Myers et al., 2005) it is not unlikely that between any two single-nucleotide polymorphisms (SNPs) there has also been some recombination. Recombination events often go undetected (Hudson and Kaplan, 1985; Song et al., 2005; Gusfield, 2014) and their effects on patterns of variation within loci can be subtle. But over longer stretches of DNA up to the chromosome scale, recombination dramatically decouples gene genealogies, revealing patterns of local selection and variation in times to common ancestry.

For a locus with infinite-sites mutation rate θ and recombination rate $\rho = 2N_e r$, where r is the probability of recombination between the two ends of the locus, Hudson (1983b) found

$$\mathrm{Var}[S] \simeq \theta \sum_{i=1}^{n-1} \frac{1}{i} + V(\rho)\theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2}$$

as an interpolation between the results of Ewens (1974) and Watterson (1975). That is, $V(\rho)$ decreases from 1 to 0 as ρ increases from zero to infinity. A simple formula for $V(\rho)$ can be found for $n = 2$ based on previous results of Griffiths (1981) about the correlation of segregating sites at two loci. Kaplan and Hudson (1985) further studied the genealogical structure of this model. Simonsen and Churchill (1997) made a detailed investigation of the Markov process of coalescence at two loci for $n = 2$. McVean (2002) showed how these results going back to Griffiths (1981) can be used to understand linkage disequilibrium.

The effects of both recombination and selection were not at all subtle in the first population survey of DNA sequence variation (Kreitman, 1983). This led to a remarkable series of papers about coalescent processes with recombination and strong selection (Kaplan et al., 1988; Hudson and Kaplan, 1988; Kaplan et al., 1989, 1991; Hudson and Kaplan, 1995). In these models, selection determines the frequencies of alleles, which define subpopulations of changing size over time, and recombination allows linked sites to 'migrate' between alleles. Barton et al. (2004) extended this framework to include weak selection. Loci close to a locus under selection have gene genealogies similar to that of the selected locus, which differ greatly from those at distant neutral loci. For example, variation will be depressed near the site of a "selective sweep" (Maynard Smith and Haigh, 1974; Kaplan et al., 1989; Barton, 1998; Durrett and Schweinsberg, 2004;

Etheridge et al., 2006). Coop and Griffiths (2004) developed an importance-sampling method to infer the timing and strength of a sweep at an infinite-sites locus without recombination. This complemented haplotype-based methods (Sabeti et al., 2002; Voight et al., 2006; Vitti et al., 2013) and SNP-based methods (Stephan et al., 1992; Kim and Stephan, 2002; Nielsen et al., 2005) of locating selective sweeps in genomes with recombination. Hermisson and Pennings (2005) and Pennings and Hermisson (2006a,b) introduced the notion of "soft sweeps" resulting from either multiple parallel mutations or standing genetic variation. Ferrer-Admetlla et al. (2014) developed a haplotype-based method of detecting and locating soft sweeps. Barton et al. (2013) studied sweeps in continuously distributed populations with migration.

Arguably the greatest contribution of Hudson (1983b) was the algorithm for simulating gene genealogies with recombination which became the program `ms` (Hudson, 1990, 2002). In `ms`, recombination breaks ancestral haplotypes apart and coalescence brings them together. Only recombination events which affect distributions of genetic ancestry are simulated, and the algorithm stops when every site has reached its most recent common ancestor. The outcome is a set of coalescent trees along the sampled sequences, with very much or very little structure in common depending on recombination. A set of such correlated trees is commonly referred to as an ancestral recombination graph, or ARG. However, it should be noted that ARG originally referred to a probability model, in which different sorts of recombination events need not be distinguished and the ancestry of the sample would be followed back beyond the most recent common ancestors to the time when all ancestral material is on a single chromosome (Griffiths, 1991; Griffiths and Marjoram, 1997). For this reason, Kelleher et al. (2016) referred to Hudson's series of linked genealogies as the 'little' ARG when they updated and reimplemented `ms` with a new data structure to make the program `msprime` for efficient simulation of large samples of chromosome-length sequences.

6.1. Moving along the chromosome, approximations and big data

Wiuf and Hein (1999) reimagined Hudson's model of coalescence with recombination as a point process along a chromosome. They showed that ARGs can be generated starting with a gene genealogy at one end of a chromosome then modeling how recombination and subsequent coalescence lead to different gene genealogies at more distant sites. This process is complicated because what happens far along the chromosome depends on everything which happened up to that point. McVean and Cardin (2005) suggested the sequentially Markov coalescent (SMC) which keeps only the immediately preceding gene genealogy as it moves along the chromosome. A modification by Marjoram and Wall (2006) gave the SMC$'$ which is a better approximation to coalescence with recombination (Chen et al., 2009; Wilton et al., 2015). Li and Durbin (2011) used the SMC in their groundbreaking Hidden Markov Model (HMM) which traverses a pair of genomes, inferring the series of coalescence times along each chromosome for inference of effective population sizes through time. Spence et al. (2018) provide a recent review of further developments of HMMs.

The urn models of successive sampling or forward-time generative models for sample distributions, which may be traced back to Ewens (1972), have also been applied to coalescence with recombination. Fearnhead and Donnelly (2001) modeled successive haplotypes as imperfect mosaics of those already sampled. Li and Stephens (2003) proposed a more efficient version of this idea and made the connection to successive sampling clear in their "product of approximate conditionals" (PAC) method, spurring much future work (Paul and Song, 2010; Steinrücken et al., 2013; Song, 2016). Rasmussen et al. (2014) added an explicit element of time to these copying models. Their program *ARGweaver* builds ('little') ARGs by modeling the ancestry of samples one at a time, connecting them probabilistically to a growing, larger ARG of the entire sample. Palacios et al. (2015) used results of one run of *ARGweaver* as input into a method of inferring past effective population sizes. Stern et al. (2019) took 20 such ARGs as input into their method of inferring allele frequency trajectories under selection at loci across the genome, and remarked that *ARGweaver* is the only currently available tool for sampling ARGs in proportion to their probabilities given the data.

These new approaches to recombination have facilitated recent applications to humans. The technical issues facing human population geneticists today are largely about scaling up to analyze enormous data sets in resources such as the Simons Genome Diversity Project (Mallick et al., 2016), the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015) and the UK Biobank (Bycroft et al., 2018). The current state of the art is illustrated by Kelleher et al. (2019) and Speidel et al. (2019) who reconstructed time-ordered series of mutations and coalescent events at loci along each human chromosome, and by Albers and McVean (2020) who estimated the ages of mutations at millions of SNPs across the human genome.

All three of these works use the standard neutral coalescent model with recombination and infinite-sites mutation as a prior model. Kelleher et al. (2019) and Speidel et al. (2019) applied the techniques of Li and Stephens (2003), denoted L&S below. Using a minimal representation of partially shared trees along a chromosome, Kelleher et al. (2019) first built a set of candidate ancestral haplotypes, assuming that more-frequent mutations are older than less-frequent ones, then adapted L&S to infer how these haplotypes are related and how present-day sequences are descended from them via recombination. Speidel et al. (2019) customized L&S to estimate mutational divergences between all pairs of samples at select sites along each chromosome. They applied a deterministic algorithm to construct series of rooted trees from these pairwise divergence matrices, invoking recombination when required under the infinite-sites model, then mapped mutations corresponding to each polymorphic site onto the branches of the trees. Albers and McVean (2020) took a composite likelihood approach using pairs of chromosomes to estimate the ages of mutations. If just one of the pair has a particular mutation, then the most recent common ancestor for that pair at that site must be older than the mutation. If both chromosomes have it, then their most recent common ancestor must be younger than the mutation. Albers and McVean (2020) tailored a sequentially Markov coalescent process using both recombination and (other) mutations to refine these estimates of pairwise times to common ancestry, effectively bounding the age of each mutation from above and below.

These three new works use approximations and make bold, though well-reasoned, decisions in order to infer human gene genealogies and ages of mutations on a grand scale. They have created the first iteration of a new global resource (Harris, 2019; Albers and McVean, 2020). The idea of performing detailed coalescent-based analyses with mutation and recombination on this scale, for example along the lines of what is done in MCMC programs such as LAMARC (Kuhner, 2006), hardly seems worth considering today. At the same time, it will be important to understand the types of errors incurred in approximate methods, or partly deterministic methods such as those of Kelleher et al. (2019) and Speidel et al. (2019), and to measure the magnitudes of these errors, so further improvements can be made. One might predict that the ancient and the recent past will require different treatments. For the ancient past, which may never be strongly constrained by data, modeling uncertainty in haplotype structures seems vital. For the recent past, it may be desirable to account explicitly for organismal genealogies, or population pedigrees (Wakeley et al., 2012; Wilton et al., 2017; Ko and Nielsen, 2019; Kelleher et al., 2018).

7. Closing remarks

This Commentary has highlighted key contributions to the theory of gene genealogies which have appeared in *TPB*, and has sketched some further developments up to today. It has focused on infinite-alleles and infinite-sites mutation, and on models of coalescence with recombination which underpin current applications. Well-mixed populations and selective neutrality have provided the context for describing and understanding patterns of genetic variation. Extensions to deal with population structure have been mentioned mostly in passing. Besides the large body of work on structured coalescent processes with migration, which Nordborg (2001) and others have reviewed, detailed treatments of gene trees versus species trees (Rosenberg, 2002; Mehta and Rosenberg, 2019), admixture or introgression (Buzbas and Verdu, 2018; Soraggi and Wiuf, 2019), the ages of mutations (Wiuf and Donnelly, 1999; Stephens, 2000) and a range of other topics have appeared in these pages. A line of work on selection, which is mostly of theoretical interest but deserves to be mentioned because it has largely taken place in *TPB*, is the unfolding of ideas about the ancestral selection graph (Krone and Neuhauser, 1997; Neuhauser, 1999; Slade, 2000a,b; Mano, 2009; Kluth and Baake, 2013; Pokalyuk and Pfaffelhuber, 2013; Lenz et al., 2015). It may be hoped that the next fifty years of *TPB* will be equally fruitful in the application of mathematics and probabilistic thinking to population biology.

Acknowledgments

I thank Noah Rosenberg and two anonymous reviewers for detailed comments which greatly improved the work.

References

Albers, P.K., McVean, G., 2020. Dating genomic variants and shared ancestry in population-scale sequencing data. PLoS Biol. 18, e3000586.

Arratia, R., Barbour, A.D., Tavaré, S., 2003. Logarithmic Combinatorial Structures: A Probabilistic Approach. In: EMS Monographs in Mathematics, European Mathematical Society.

Barton, N.H., 1998. The effect of hitch-hiking on neutral genealogies. Genet. Res. Camb. 72, 123–133.

Barton, N.H., Etheridge, A.M., Kelleher, J., Véber, A., 2013. Genetic hitchhiking in spatially extended populations. Theor. Popul. Biol. 87, 75–89.

Barton, N.H., Etheridge, A.M., Sturm, A.K., 2004. Coalescence in a random background. Ann. Appl. Probab. 14, 754–785.

Berestycki, N., 2009. Recent progress in coalescent theory. Ensaios Mat. 16, 1–193.

Birkner, M., Blath, J., Capaldo, M., Etheridge, A., Möhle, M., Schweinsberg, J., Wakolbinger, A., 2005. Alpha-stable branching processes and beta-coalescents. Electron. J. Probab. 10, 303–325.

Brown, W.M., 1980. Polymorphism in mitochondrial DNA of humans revealed by restriction endonuclease analysis. Proc. Natl. Acad. Sci. USA 70, 3605–3609.

Burden, C.J., Tang, Y., 2017. Rate matrix estimation from site frequency data. Theor. Popul. Biol. 113, 23–33.

Buzbas, E.O., Verdu, P., 2018. Inference on admixture fractions in a mechanistic model of recurrent admixture. Theor. Popul. Biol. 122, 149–157.

Bycroft, C., et al., 2018. The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209.

Chen, G.K., Marjoram, P., Wall, J.D., 2009. Fast and flexible simulation of DNA sequence data. Genome Res. 19, 136–142.

Coop, G., Griffiths, R.C., 2004. Ancestral inference on gene trees under selection. Theor. Popul. Biol. 66, 219–232.

Donnelly, P., 1986. Partition structures, Polya urns, the Ewens sampling formula, and the ages of alleles. Theor. Popul. Biol. 30, 271–288.

Donnelly, P., Kurtz, T.G., 1996a. The asymptotic behavior of an urn model arising in population genetics. Stochastic Process. Appl. 64, 1–16.

Donnelly, P., Kurtz, T.G., 1996b. A countable representation of the Fleming-Viot measure-valued diffusion. Ann. Probab. 24, 698–742.

Donnelly, P., Kurtz, T.G., 1999. Particle representations for measure-valued population models. Ann. Probab. 27, 166–205.

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

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29, 1969–1973.

Dung, S.K., López, A., Barragan, E.L., Reyes, R.-J., Thu, R., Castellanos, E., Catalan, F., Huerta-Sánchez, E., Rohlfs, R.V., 2019. Illuminating women's hidden contribution to historical theoretical population genetics. Genetics 211, 363–366.

Durrett, R., 2008. Probability Models for DNA Sequence Evolution, second ed. Springer, New York.

Durrett, R., Schweinsberg, J., 2004. Approximating selective sweeps. Theor. Popul. Biol. 66, 129–138.

Etheridge, A.M., Kurtz, T.G., 2019. Genealogical constructions of population models. Ann. Probab. 47, 1827–1910.

Etheridge, A.M., Pfaffelhuber, P., Wakolbinger, A., 2006. An approximate sampling formula under genetic hitchhiking. Ann. Appl. Probab. 16, 685–729.

Ethier, S.N., Griffiths, R.C., 1987. The infinitely-many-sites model as a measure valued diffusion. Ann. Probab. 15, 515–545.

Evans, S.N., Shvets, Y., Slatkin, M., 2007. Non-equilibrium theory of the allele frequency spectrum. Theor. Popul. Biol. 71, 109–119.

Ewens, W.J., 1972. The sampling theory of selectively neutral alleles. Theor. Popul. Biol. 3, 87–112.

Ewens, W.J., 1974. A note on the sampling theory for infinite alleles and infinite sites models. Theor. Popul. Biol. 6, 143–148.

Ewens, W.J., 1990. Population genetics theory – the past and the future. In: Lessard, S. (Ed.), Mathematical and Statistical Developments of Evolutionary Theory. Kluwer Academic Publishers, Amsterdam, pp. 177–227.

Ewens, W.J., 2004. Mathematical Population Genetics, Volume I: Theoretical Foundations. Springer-Verlag, Berlin.

Fearnhead, P., Donnelly, P., 2001. Estimating recombination rates from population genetic data. Genetics 159, 1299–1318.

Felsenstein, J., 1971. The rate of loss of multiple alleles in finite haploid populations. Theor. Popul. Biol. 2, 391–405.

Felsenstein, J., Kuhner, M.K., Yamato, J., Beerli, P., 1999. Likelihoods on coalescents: A Monte Carlo sampling approach to inferring parameters from population samples of molecular data. In: Seillier-Moiseiwitsch, F. (Ed.), Statistics in Molecular Biology and Genetics. In: IMS Lecture Notes-Monograph Series, vol. 33, Institute of Mathematical Statistics, Hayward, California, pp. 163–185.

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., Nielsen, R., 2014. On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol. Biol. Evol. 31, 1275–1291.

Ferretti, L., Klassmann, A., Raineri, E., Ramos-Onsins, S.E., Wiehe, T., Achaz, G., 2018. The neutral frequency spectrum of linked sites. Theor. Popul. Biol. 123, 70–79.

Fisher, R.A., 1930. The Genetical Theory of Natural Selection. Clarendon, Oxford.

Fu, Y.-X., 1995. Statistical properties of segregating sites. Theor. Popul. Biol. 48, 172–197.

Fuerst, P.A., Ferrell, R.E., 1980. The stepwise mutation model: an experimental evaluation utilizing hemoglobin variants. Genetics 94, 185–201.

Goldstein, D.B., Linares, A.R., Feldman, M.W., Cavalli Sforza, L.L., 1995. An evaluation of genetic distances for use with microsatellite loci. Genetics 139, 463–471.

Griffiths, R.C., 1979. Exact sampling distributions from the infinite neutral alleles model. Adv. Appl. Probab. 11, 326–354.

Griffiths, R.C., 1980. Lines of descent in the diffusion approximation of neutral wright-fisher models. Theor. Popul. Biol. 17, 37–50.

Griffiths, R.C., 1981. Neutral two-locus multiple allele models with recombination. Theor. Popul. Biol. 19, 169–186.

Griffiths, R.C., 1991. The two-locus ancestral graph. In: Basawa, I.V., Taylor, R.L. (Eds.), Selected Proceedings of the Symposium on Applied Probability. Institute of Mathematical Statistics, Hayward, CA, USA, pp. 100–117.

Griffiths, R.C., Marjoram, P., 1997. An ancestral recombination graph. In: Donnelly, P., Tavaré, S. (Eds.), Progress in Population Genetics and Human Evolution. In: IMA Volumes in Mathematics and Its Applications, vol. 87, Springer-Verlag, New York, pp. 257–270.

Griffiths, R.C., Tavaré, S., 1994a. Ancestral inference in population genetics. Statist. Sci. 9, 307–319.

Griffiths, R.C., Tavaré, S., 1994b. Simulating probability distributions in the coalescent. Theor. Popul. Biol. 46, 131–159.

Griffiths, R.C., Tavaré, S., 1998. The age of a mutation in a general coalescent tree. Commun. Statist. – Stoch. Models 14, 273–295.

Gusfield, D., 2014. ReCombinatorics: the Algorithms of Ancestral Recombination Graphs and Explicit Phylogenetic Networks. MIT Press, Cambridge, Massachusetts.

Harris, H., 1966. Enzyme polymorphism in man. Proc. R. Soc. London, Ser. B 164, 298–310.

Please cite this article as: J. Wakeley, Developments in coalescent theory from single loci to chromosomes. Theoretical Population Biology (2020), https://doi.org/10.1016/j.tpb.2020.02.002.

Harris, K., 2019. From a database of genomes to a forest of evolutionary trees. Nat. Genet. 51, 1304–1307.

Hein, J., Schierup, M.H., Wiuf, C., 2005. Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory. Oxford University Press, Oxford.

Hermisson, J., Pennings, P.S., 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169, 2335–2352.

Hoppe, F.M., 1984. Pó1ya-like urns and the Ewens' sampling formula. J. Math. Biol. 20, 91–94.

Hoppe, F.M., 1987. The sampling theory of neutral alleles and an urn model in population genetics. J. Math. Biol. 25, 123–159.

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

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

Hudson, R.R., 1990. Gene genealogies and the coalescent process. In: Futuyma, D.J., Antonovics, J. (Eds.), Oxford Surveys in Evolutionary Biology, Vol. 7. Oxford University Press, Oxford, pp. 1–44.

Hudson, R.R., 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18, 337–338.

Hudson, R.R., Kaplan, N.L., 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111, 147–164.

Hudson, R.R., Kaplan, N.L., 1988. The coalescent process in models with selection and recombination. Genetics 120, 831–840.

Hudson, R.R., Kaplan, N.L., 1995. Deleterious background selection with recombination. Genetics 141, 1605–1617.

Jenkins, P.A., Mueller, J.W., Song, Y.S., 2014. General triallelic frequency spectrum under demographic models with variable population size. Genetics 196, 295–311.

Johnson, N.L., Kotz, S., 1969. Distributions in Statistics: Discrete Distributions. Houghton Mifflin, Boston.

Kaplan, N.L., Darden, T., Hudson, R.R., 1988. Coalescent process in models with selection. Genetics 120, 819–829.

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

Kaplan, N.L., Hudson, R.R., Iizuka, M., 1991. Coalescent processes in models with selection, recombination and geographic subdivision. Genet. Res., Camb. 57, 83–91.

Kaplan, N.L., Hudson, R.R., Langley, C.H., 1989. The "hitchhiking effect" revisited. Genetics 123, 887–899.

Karlin, S., McGregor, J., 1972. Addendum to a paper of W. Ewens. Theor. Popul. Biol. 3, 113–116.

Kelleher, J., Etheridge, A.M., McVean, G., 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput. Biol. 12, e10048427.

Kelleher, J., Thornton, K.R., Ashander, J., Ralph, P.L., 2018. Efficient pedigree recording for fast population genetics simulation. PLoS Comput. Biol. 14, e1006581.

Kelleher, J., Wong, Y., Wohns, A.W., Fadil, C., Albers, P.K., McVean, G., 2019. Inferring whole-genome histories in large population datasets. Nat. Genet. 51, 1330–1338.

Kelly, F.P., 1979. Reversibility and stochastic networks. John Wiley & Sons, New York.

Kim, Y., Stephan, W., 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160, 765–777.

Kimura, M., 1955. Random genetic drift in a multi-allelic locus. Evolution 9, 419–435.

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., 1971. Theoretical foundation of population genetics at the molecular level. Theor. Popul. Biol. 2, 174–208.

Kimura, M., Crow, J.F., 1964. The number of alleles that can be maintained in a finite population. Genetics 49, 725–738.

Kingman, J.F.C., 1975. Random discrete distributions. J. R. Stat. Soc. Ser. B Stat. Methodol. 37, 1–15.

Kingman, J.F.C., 1976. Coherent random walks arising in some genetical models. Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 351, 19–31.

Kingman, J.F.C., 1977a. A note on multidimensional models of neutral mutation. Theor. Popul. Biol. 11, 285–290.

Kingman, J.F.C., 1977b. The population structure associated with the Ewens sampling formula. Theor. Popul. Biol. 11, 274–283.

Kingman, J.F.C., 1982a. The coalescent. Stochastic Process. Appl. 13, 235–248.

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

Kingman, J.F.C., 1982c. On the genealogy of large populations. J. Appl. Probab. 19A, 27–43.

Kingman, J.F.C., 2000. Origins of the coalescent: 1974–1982. Genetics 156, 1461–1463.

Kluth, S., Baake, E., 2013. The moran model with selection: Fixation probabilities, ancestral lines, and an alternative particle representation. Theor. Popul. Biol. 90, 104–112.

Ko, A., Nielsen, R., 2019. Joint estimation of pedigrees and effective population size using Markov chain Monte Carlo. Genetics 212, 855–868.

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

Krone, S.M., Neuhauser, C., 1997. Ancestral processes with selection. Theor. Popul. Biol. 51, 210–237.

Kuhner, M.K., 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 22, 768–770.

Kuhner, M.K., Yamato, J., Felsenstein, J., 1995. Estimating effective population size and mutation rate from sequence data using Metropolois-Hastings sampling. Genetics 140, 1421–1430.

Leffler, E.M., Bullaughey, K., Matute, D.R., Meyer, W.K., Ségurel, L., Venkat, A., Andolfatto, P., Przeworski, M., 2012. Revisiting an old riddle: What determines genetic diversity levels within species?. PLoS Biol. 10(9), e1001388.

Lenz, U., Kluth, S., Baake, E., Wakolbinger, A., 2015. Looking down in the ancestral selection graph: A probabilistic approach to the common ancestor type distribution. Theor. Popul. Biol. 103, 27–37.

Lewontin, R.C., Hubby, J.L., 1966. A molecular approach 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.

Li, H., Durbin, R., 2011. Inference of population history from individual whole-genome sequences. Nature 475, 493–496.

Li, N., Stephens, M., 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165, 2213–2233.

Lichten, M., Goldman, A.S.H., 1995. Meiotic recombination hotspots. Annu. Rev. Genet. 29, 423–444.

Lynch, M., 2007. The Origins of Genome Architecture. Sinauer Associates, Inc., Sunderland, Massachusetts.

Malécot, G., 1941. Etude mathématique des populations Mendélienne. Ann. Univ. Lyon Sci. Sec. A 4, 45–60.

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

Mallick, S., et al., 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538, 201–206.

Mano, S., 2009. Duality, ancestral and diffusion processes in models with selection. Theor. Popul. Biol. 75, 164–175.

Marjoram, P., Wall, J.D., 2006. Fast "coalescent" simulation. BMC Genet. 7, 16.

Maynard Smith, J., Haigh, J., 1974. The hitchhiking effect of a favorable gene. Genet. Res. 23, 23–35.

McVean, G.A.T., 2002. A genealogical interpretation of linkage disequilibrium. Genetics 162, 987–991.

McVean, G.A.T., Cardin, N.J., 2005. Approximating the coalescent with recombination. Philos. Trans. R. Soc. B 360, 1387–1393.

Mehta, R.S., Rosenberg, N.A., 2019. The probability of reciprocal monophyly of gene lineages in three and four species. Theor. Popul. Biol. 129, 133–147.

Möhle, M., 1998. Robustness results for the coalescent. J. Appl. Probab. 35, 438–447.

Möhle, M., 2001. Forward and backward diffusion approximations for haploid exchangeable population models. Stochastic Process. Appl. 95, 133–149.

Möhle, M., Sagitov, S., 2001. A classification of coalescent processes for haploid exchangeable population models. Ann. Probab. 29, 1547–1562.

Moran, P.A.P., 1958. Random processes in genetics. Proc. Camb. Phil. Soc. 54, 60–71.

Moran, P.A.P., 1975. Wandering distributions and the electrophoretic profile. Theor. Popul. Biol. 8, 318–330.

Moran, P.A.P., 1976. Wandering distributions and the electrophoretic profile II. Theor. Popul. Biol. 10, 145–149.

Myers, S., Bottolo, L., Freeman, C., McVean, G., Donnelly, P., 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310, 321–324.

Neuhauser, C., 1999. The ancestral graph and gene genealogy under frequency-dependent selection. Theor. Popul. Biol. 56, 203–214.

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G., Bustamante, C., 2005. Genomic scans for selective sweeps using SNP data. Genome Res. 15, 1566–1575.

Nordborg, M., 2001. Coalescent theory. In: Balding, D.J., Bishop, M.J., Cannings, C. (Eds.), Handbook of Statistical Genetics. John Wiley & Sons, Chichester, England, pp. 179–212.

Ohta, T., Kimura, M., 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res., Camb. 22, 201–204.

Palacios, J.A., Wakeley, J., Ramachandran, S., 2015. Bayesian nonparametric inference of population size changes from sequential genealogies. Genetics 201, 281–304.

Paul, J.S., Song, Y.S., 2010. A principled approach to deriving approximate conditional sampling distributions in population genetics models with recombination. Genetics 186, 321–328.

Pennings, P.S., Hermisson, J., 2006a. Soft sweeps II: Molecular population genetics of adaptation from recurrent mutation or migration. Mol. Biol. Evol. 23, 1076–1084.

Pennings, P.S., Hermisson, J., 2006b. Soft sweeps III: The signature of positive selection from recurrent mutation. PLoS Genet. 2 (12), e186.

Pitman, J., 1999. Coalescents with multiple collisions. Ann. Probab. 27, 1870–1902.

Pokalyuk, C., Pfaffelhuber, P., 2013. The ancestral selection graph under strong directional selection. Theor. Popul. Biol. 87, 25–33.

Polanski, A., Bobrowski, A., Kimmel, M., 2003. A note on distributions of times to coalescence under time-dependent population size. Theor. Popul. Biol. 63, 33–40.

Polanski, A., Kimmel, M., 2003. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics 165, 427–436.

Ramshaw, J.A.M., Coyne, J.A., Lewontin, R.C., 1979. The sensitivity of gel electrophoresis as a detector of genetic variation. Genetics 93, 1019–1037.

Rasmussen, M.D., Gronau, M.J.H.I., Siepel, A., 2014. Genome-wide inference of ancestral recombination graphs. PLoS Genet. 10, e1004342.

Rosenberg, N.A., 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61, 225–247.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., et al., 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419, 832–837.

Sagitov, S., 1999. The general coalescent with asynchronous mergers of ancestral lines. J. Appl. Probab. 36, 1116–1125.

Sainudiin, R., Stadler, T., Véber, A., 2015. Finding the best resolution for the Kingman-Tajima coalescent: theory and applications. J. Math. Biol. 70, 1207–1247.

Sainudiin, R., Véber, A., 2018. Full likelihood inference from the site frequency spectrum based on the optimal tree resolution. Theor. Popul. Biol. 124, 1–15.

Sargsyan, O., 2015. An analytical framework in the general coalescent tree setting for analyzing polymorphisms created by two mutations. J. Math. Biol. 70, 913–956.

Schweinsberg, J., 2000. Coalescents with simultaneous multiple collisions. Electron. J. Probab. 5, 1–50.

Shah, D.M., Langley, C.H., 1979. Inter- and intraspecific variation in restriction maps of *Drosophila* mitochondrial DNAs. Nature 281, 696–699.

Simonsen, K.L., Churchill, G.A., 1997. A Markov chain model of coalescence with recombination. Theor. Popul. Biol. 52, 43–59.

Sjödin, P., Kaj, I., Krone, S., Lascoux, M., Nordborg, M., 2005. On the meaning and existence of an effective population size. Genetics 169, 1061–1070.

Slade, P.F., 2000a. Most recent common ancestor distributions in genealogies under selection. Theor. Popul. Biol. 58, 291–305.

Slade, P.F., 2000b. Simulation of selected genealogies. Theor. Popul. Biol. 57, 35–49.

Slatkin, M., 1995. A measure or population subdivision based on microsatellite allele frequencies. Genetics 139, 457–462.

Song, Y.S., 2016. Na Li and Matthew Stephens on modeling linkage disequilibrium. Genetics 203, 1005–1006.

Song, Y.S., Wu, Y., Gusfield, D., 2005. Efficient computation of close lower and upper bounds on the minimum number of recombinations in biological sequence evolution. Bioinformatics 21, 413–422.

Soraggi, S., Wiuf, C., 2019. General theory for stochastic admixture graphs and f-statistics. Theoret. Popu. Biol. 125, 56–66.

Speidel, L., Forest, M., Sinan, S., Myers, S.R., 2019. A method for genome-wide genealogy estimation for thousands of samples. Nat. Genet. 51, 1321–1329.

Spence, J.P., Steinrücken, M., Terhorst, J., Song, Y.S., 2018. Inference of population history using coalescent HMMs: review and outlook. Curr. Opin. Genet. Dev. 53, 70–76.

Steinrücken, M., Paul, J.S., Song, Y.S., 2013. A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. Theor. Popul. Biol. 87, 51–61.

Stephan, W., Wiehe, T.H.E., Lenz, M.W., 1992. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor. Popul. Biol. 41, 237–254.

Stephens, M., 2000. Times on trees and the age of an allele. Theor. Popul. Biol. 57, 109–119.

Stephens, M., Donnelly, P., 2000. Inference in molecular population genetics. J. R. Stat. Soc. Ser. B 62, 605–655.

Stern, A.J., Wilton, P.R., Nielsen, R., 2019. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. PLoS Genet. 15, e1008384.

Strobeck, C., 1983. Estimation of the neutral mutation rate in a finite population from DNA sequence data. Theor. Popul. Biol. 24, 160–172.

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

Takahata, N., 1988. The coalescent in two partially isolated diffusion populations. Genet. Res., Camb. 53, 213–222.

Tavaré, S., 1984. Lines-of-descent and genealogical processes, and their application in population genetic models. Theor. Popul. Biol. 26, 119–164.

Tavaré, S., 2004. Ancestral inference in population genetics. In: Cantoni, O., Tavaré, S., Zeitouni, O. (Eds.), École d'Été de Probabilités de Saint-Flour XXXI – 2001. In: Lecture Notes in Mathematics, vol. 1837, Springer-Verlag, Berlin, pp. 1–188.

The 1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. Nature 526, 68–74.

Tian, X., Browning, B.L., Browning, S.R., 2019. Estimating the genome-wide mutation rate with three-way identity by descent. Am. J. Hum. Genet. 105, 883–893.

Valdes, A.M., Slatkin, M., Freimer, N.B., 1993. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics 133, 737–749.

Vitti, J.J., Grossman, S.R., Sabeti, P.C., 2013. Detecting natural selection in genomic data. Annu. Rev. Genet. 47, 97–120.

Voight, B.F., Kudaravalli, S., Wen, X., Pritchard, J.K., 2006. A map of recent positive selection in the human genome. PLoS Biol. 4, e72.

Wakeley, J., 2008. Coalescent Theory: An Introduction. Roberts & Company Publishers, Greenwood Village, Colorado.

Wakeley, J., King, L., Low, B.S., Ramachandran, S., 2012. Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. Genetics 190, 1433–1445.

Wakeley, J., Sargsyan, O., 2009. Extensions of the coalescent effective population size. Genetics 181, 341–345.

Watterson, G.A., 1974. The sampling theory of selectively neutral alleles. Adv. Appl. Probab. 6, 463–488.

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

Watterson, G.A., 1976a. Reversibility and the age of an allele. I. Moran's infinitely many neutral alleles model. Theor. Popul. Biol. 10, 239–253.

Watterson, G.A., 1976b. The stationary distribution of the infinitely many neutral alleles diffusion model. J. Appl. Probab. 13, 639–651.

Watterson, G.A., 1977. Heterosis or neutrality? Genetics 85, 789–814.

Watterson, G.A., 1978. The homozygosity test of neutrality. Genetics 88, 405–417.

Watterson, G.A., 1984. Lines of descent and the coalescent. Theor. Popul. Biol. 26, 77–92.

Watterson, G.A., Guess, H.A., 1977. Is the most frequent allele the oldest? Theor. Popul. Biol. 11, 141–160.

Wilton, P.R., Baduel, P., Landon, M.M., Wakeley, J., 2017. Population structure and coalescence in pedigrees: Comparisons to the structured coalescent and a framework for inference. Theor. Popul. Biol. 115, 1–12.

Wilton, P.R., Carmi, S., Hobolth, A., 2015. The SMC′ is a highly accurate approximation to the ancestral recombination graph. Genetics 200, 343–355.

Wiuf, C., Donnelly, P., 1999. Conditional genealogies and the age of a neutral mutant. Theor. Popul. Biol. 56, 183–201.

Wiuf, C., Hein, J., 1999. Recombination as a point process along sequences. Theor. Popul. Biol. 55, 248–259.

Wright, S., 1922. Coefficients of inbreeding and relationship. Am. Nat. 56, 330–338.

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

Wu, Y., 2010. Exact computation of coalescent likelihood for panmictic and subdivided populations under the infinite sites model. IEEE/ACM Trans. Comput. Biol. Bioinform. 7, 611–618.