

Coalescent Entanglement and the Conditional Dependence of the Times to Common Ancestry of Mutually Exclusive Pairs of Individuals

STEVEN WU, KATIA KOELLE, AND ALLEN RODRIGO

From the Department of Biology, Duke University, Box 90338, Durham, NC 27708 (Wu, Koelle, and Rodrigo); and The National Evolutionary Synthesis Center, 2024W. Main St., Suite A200, Durham, NC 27705 (Rodrigo).

Address correspondence to Allen Rodrigo at the address above, or e-mail: a.rodrigo@nescent.org.

Abstract

The Kingman coalescent is a continuous-time diffusion approximation of the times to common ancestry of a sample of individuals drawn from a Wright–Fisher population. Here, we use the coalescent to answer a simple question: if we know the ancestry of 2 randomly sampled individuals in the population, what does it tell us about the ancestry of 2 other randomly sampled individuals? We show that there is a conditional dependency between the times to common ancestry between pairs of randomly sampled individuals. We call this “coalescent entanglement,” and we demonstrate its effects through simulation. The effects of entanglement extend beyond the coalescent to phylogenetic birth–death processes in general. Entanglement also exerts its effects when the pairs of individuals chosen share no common lineages in the paths that connect the individuals in each pair.

Key words: *coalescent, genealogy, phylogeny, entanglement*

Genealogies, as most people know, depict the relationships amongst individuals via the ancestors that these individuals share. The times that separate individuals from their common ancestors, in turn, tell us something about the processes that have shaped the populations from which these individuals are sampled. For instance, if one randomly samples 2 individuals from a large population, one expects that a far greater number of generations separate them from their common ancestor, than if these 2 individuals came from a small population. Intuitively, there should be a relationship between population size and the time it takes for lineages on a genealogy to meet, or “coalesce,” as one moves from individuals sampled at the present time, to past ancestors. In 1982, John Kingman formalized this relationship between genealogies and population size, and introduced the coalescent to population genetics (Kingman 1982). The coalescent, as it was originally conceived, is a mathematical description of the genealogy of a sample of individuals from a large Wright–Fisher population. With the parallel expansion of DNA sequencing and computational technologies in the 80s and 90s, the coalescent has come into its own in population genetic inference and now forms the foundation on which most of our conceptual and methodological developments are built.

The coalescent is a continuous-time diffusion approximation of the distributions of times to common ancestry of the lineages of individuals sampled from a population, as one moves from the present to the past along the genealogy of these individuals. For a constant-sized Wright–Fisher population, these times are exponential random variables, with expectations that are a function of the size of the population, and the number of lineages that have yet to coalesce. Arguably, one of the most significant aspects of the coalescent is its utility as a heuristic: knowing something about the structure of genealogies provides us with the conceptual tools to think about how evolutionary phenomena as diverse as migration (Nath and Griffiths 1993; Wakeley and Lessard 2006), selection (Kaplan et al. 1988; O’Fallon et al. 2010), recombination (Hudson and Kaplan 1988; McVean and Cardin 2005), and speciation (Hobolth et al. 2007; Kondo et al. 2008; Morlon et al. 2010) have an impact on biological identity and variation.

Here, we use the coalescent for just this purpose—to investigate an apparently simple problem that happens to have a counterintuitive solution. The question is this: if you know the ancestry of 2 of your neighbors, what does it tell you about the ancestry of you and your spouse? Formally, we can restate the problem as follows: what is the

conditional expectation of the time to the most recent common ancestor (MRCA) of 2 randomly sampled individuals C and D ($t\text{MRCA}_{\text{CD}}$), given that we already know the time to the MRCA of individuals A and B ($t\text{MRCA}_{\text{AB}}$)? If we know no more about $t\text{MRCA}_{\text{CD}}$ given $t\text{MRCA}_{\text{AB}}$, then $E[t\text{MRCA}_{\text{CD}} | t\text{MRCA}_{\text{AB}}] = E[t\text{MRCA}_{\text{CD}}]$, that is, the conditional expectation of $t\text{MRCA}_{\text{CD}}$ given $t\text{MRCA}_{\text{AB}}$ would equal the marginal expectation of $t\text{MRCA}_{\text{CD}}$. More completely, the distributions of times to common ancestry of C and D would be exactly the same as if we had not known anything about $t\text{MRCA}_{\text{AB}}$. As we show below, this equality does not hold. Instead, we demonstrate analytically and by simulation that $E[t\text{MRCA}_{\text{CD}} | t\text{MRCA}_{\text{AB}}]$ is a function of $t\text{MRCA}_{\text{AB}}$. Furthermore, we show that beyond a certain time, $E[t\text{MRCA}_{\text{CD}} | t\text{MRCA}_{\text{AB}}]$ increases linearly with $t\text{MRCA}_{\text{AB}}$.

We call this phenomenon of conditional dependency between the times to common ancestry “coalescent entanglement” to draw the analogy between this process and quantum entanglement, whereby 2 apparently independent quantum mechanical systems have persistent correlated quantum states. We recognize that the qualitative results presented here will not surprise population geneticists who already work with the coalescent. Nonetheless, the precise nature of the conditional dependencies between mutually exclusive pairs of individuals has not, to our knowledge, been characterized.

We also demonstrate that this conditional dependence is seen when the branching lineages of individuals are modeled by a birth–death process, similar to that used to model species’ divergence times on a phylogeny (in which case, a more appropriate term may be “phylogenetic entanglement”). We show that even when pairs of individuals are chosen such that they share no lineages in the paths that connect individuals within each pair, there is still a conditional dependence between the times to common ancestry.

The Model

The simplest model of the coalescent assumes a panmictic, haploid population with a constant number of individuals, N , discrete generations, and an absence of selective forces acting on the population. If we sample 2 individuals from such a population at the present time, t_0 , the probability that both will share their MRCA at time t_k , which is k generations before the present is given by the geometric distribution:

$$P(t_k) = \left(\frac{1}{N}\right) \left(1 - \frac{1}{N}\right)^{k-1}. \quad (1)$$

If n individuals are sampled (where $2 \leq n \ll N$), then the probability that there will be 1 common ancestor in the preceding generation is

$$P(t_1) = \frac{n(n-1)}{2N}, \quad (2)$$

and the probability that the first MRCA of any of these $n(n-1)/2$ pairs will be at time t_k is

$$P(t_k) = \left(\frac{n(n-1)}{2N}\right) \left(1 - \frac{n(n-1)}{2N}\right)^{k-1}. \quad (3)$$

As $N \rightarrow \infty$, we can obtain a continuous-time approximation for the probability density of t_k (Kaplan et al. 1988):

$$P(t_k) = \left(\frac{n(n-1)}{2N}\right) \exp\left(-\frac{n(n-1)}{2N}k\right) \quad (4)$$

To simulate genealogies of a sample, n , under a constant population size, N , we use the following algorithm:

1. Select two individuals at random from the number of individual lineages, n , that have yet to coalesce.
2. Obtain a coalescent interval by drawing a random variable from Equation (4), and join the two lineages drawn in Step 1, after the interval has elapsed. Set n to $n - 1$.
3. If $n = 1$, stop; otherwise, return to Step 1.

The time to common ancestry of the entire genealogy is the sum of all coalescent intervals.

Results

To illustrate the phenomenon of coalescent entanglement, we simulated 10 million genealogies of 4 individuals, A, B, C, and D. For each genealogy simulated, we plot $t\text{MRCA}_{\text{CD}}$ against $t\text{MRCA}_{\text{AB}}$ (Figure 1A). We also plot $E[t\text{MRCA}_{\text{CD}} | t\text{MRCA}_{\text{AB}}]$ against $t\text{MRCA}_{\text{AB}}$ (Figure 1B). As noted above, the time (in generations) to common ancestry between any 2 lineages is distributed as an exponential random variable with rate parameter $\binom{n}{2}/N$. It is convenient to express coalescent times in units of N , in which case, times are drawn from an exponential distribution with rate parameter $\binom{n}{2}$. Hence, under the coalescent, the time (in units of N) to common ancestry of 2 randomly sampled individuals (e.g., C and D) is exponentially distributed with rate 1. Indeed, with our simulations, the marginal distribution of $t\text{MRCA}_{\text{CD}}$, taken over all values of $t\text{MRCA}_{\text{AB}}$, recovers this expected exponential distribution, as does the marginal expectation of $t\text{MRCA}_{\text{AB}}$ (Figure 1A). This is important: in the absence of any knowledge about $t\text{MRCA}_{\text{AB}}$, the distribution of $t\text{MRCA}_{\text{CD}}$ is no different from what we would expect under the standard coalescent model. For any given $t\text{MRCA}_{\text{AB}}$, however, $E[t\text{MRCA}_{\text{CD}} | t\text{MRCA}_{\text{AB}}]$ varies curvilinearly with $t\text{MRCA}_{\text{AB}}$ (Figure 1B); the curve has a minimum at approximately $0.75N$, and then tends to increase linearly.

With a proportion of simulated genealogies, the $t\text{MRCA}_{\text{CD}}$ is exactly equal to $t\text{MRCA}_{\text{AB}}$. This accounts for the values that fall on the diagonal in Figure 1A. How is this possible? The answer lies with the ordered histories of A, B, C, and D (Figure 2). An ordered history is a topology in which both the pattern of branching and the order in which the branchings occur uniquely identify the topology. In Figure 2, we have grouped the 18 possible ordered histories (obtainable with

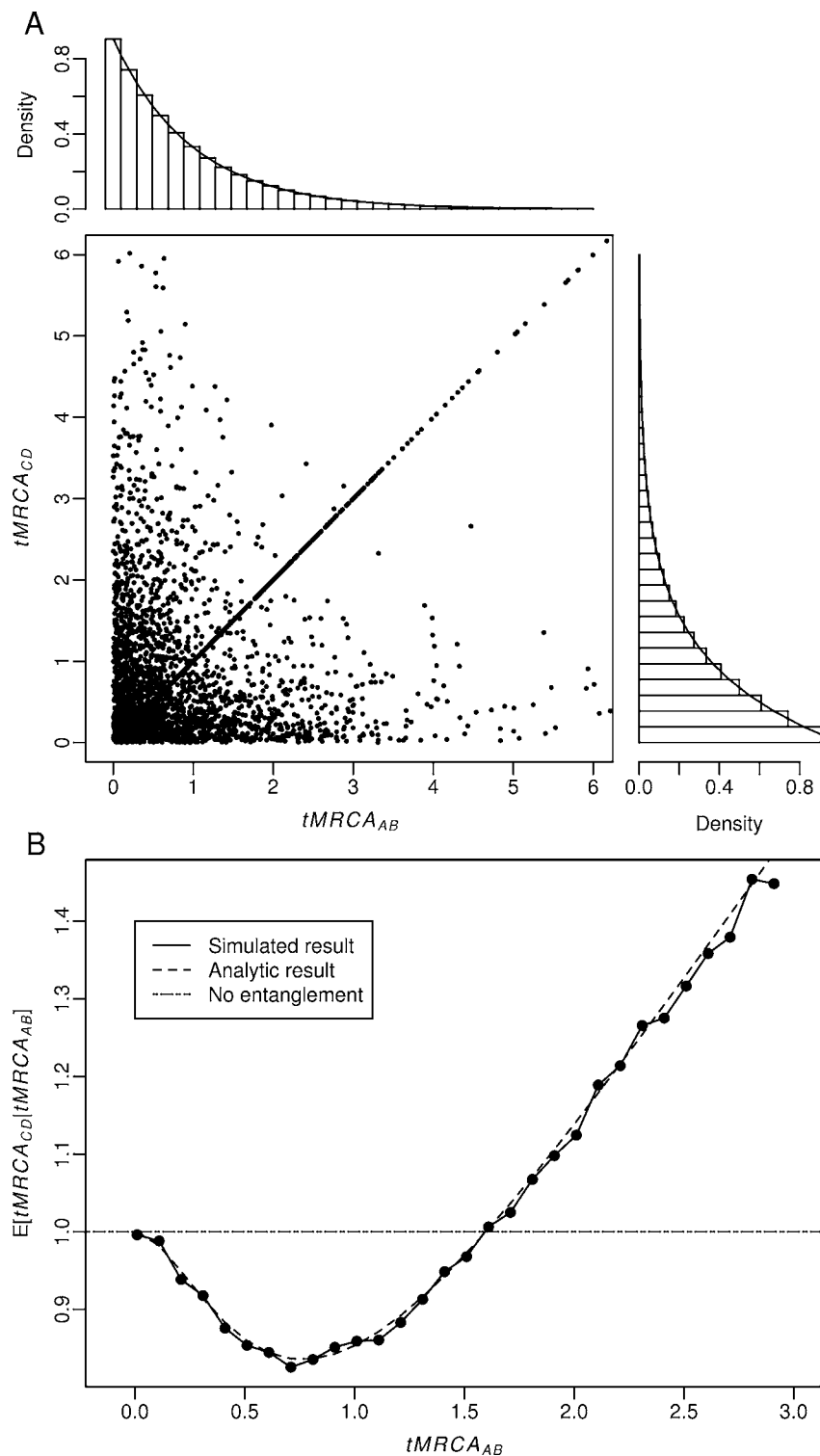


Figure 1. Plots of (A) $tMRCA_{CD}$ and (B) $E[tMRCA_{CD} | tMRCA_{AB}]$ against $tMRCA_{AB}$. Here, 10^6 simulations were performed using R, in which genealogies of A, B, C, and D were randomly generated under a constant-sized coalescent model. Time is scaled in units of N generations. In (A), the estimated densities of $tMRCA_{AB}$ and $tMRCA_{CD}$ are plotted as frequency histograms above and to the right of the scatterplot, respectively; black lines represent the marginal exponential distribution with rate parameter 1, as expected. In (B), to estimate $E[tMRCA_{CD} | tMRCA_{AB}]$ for a given $tMRCA_{AB}$, all values of $tMRCA_{CD}$ obtained in the interval $tMRCA_{AB} \pm 0.001$ were averaged. The dashed line shows the analytical results for $E[tMRCA_{CD} | tMRCA_{AB}]$ derived in the [Supplementary Material online](#).

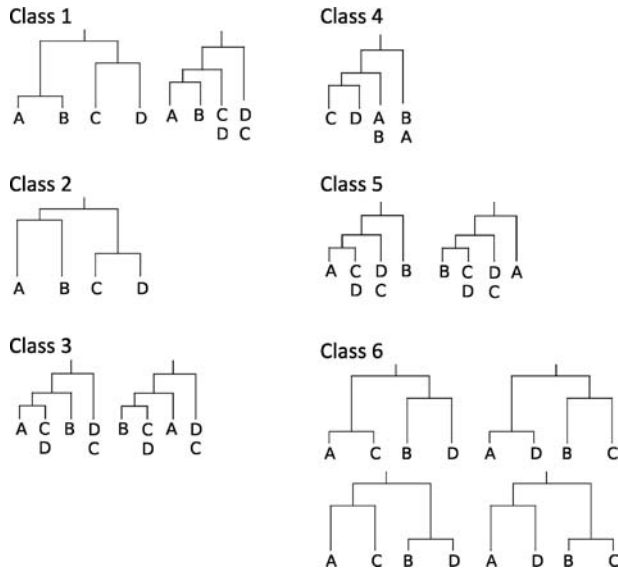


Figure 2. All possible ordered histories for A, B, C, and D. These have been grouped into classes, depending on how A and B share common lineages with C and D. Frequencies of these classes, as a function of $tMRCA_{AB}$, are given in Table 1 and can be computed analytically from the derivations presented in the [Supplementary Material online](#).

4 individuals) on the association of A and B with C and D, as well as the timing of the (A, B) coalescence. In particular, with ordered histories in Class 6, it is apparent that $tMRCA_{AB}$ must necessarily equal $tMRCA_{CD}$, because both (A, B) and (C, D) coalesce at the same point, that is, the root of the tree. This explains the simulated genealogies that fall on the diagonal in Figure 1A.

However, other ordered histories also constrain $tMRCA_{CD}$. For instance, lineages C and D coalesce before A and B in ordered histories belonging to Classes 2 and 4; hence, $tMRCA_{CD}$ must necessarily be less than $tMRCA_{AB}$. Conversely, in Class 1, $tMRCA_{CD}$ must be greater than $tMRCA_{AB}$. Interestingly, the relative proportions of these classes of ordered histories change as $tMRCA_{AB}$ increases (Table 1 and [Supplementary Material online](#)). When $tMRCA_{AB}$ is small, the most common ordered histories are in Class 1, where A and B coalesce before C and D. This is because there is only a very small probability that there will be a coalescent event earlier than the one between A and B. In contrast, when $tMRCA_{AB}$ is large, it is unlikely that there will be a coalescent event that occurs after A and B have coalesced, in which case, histories in Classes 4 through 6 dominate. In short, the fact that A, B, C, and D are part of the same evolving system imposes the conditional dependence of ancestral times. By taking into consideration these changes in the frequencies of ordered histories, the analytic formula for $E[tMRCA_{CD} | tMRCA_{AB}]$ can be derived ([Supplementary Material online](#)). We plot this analytical relationship (dashed line) in Figure 1B alongside the mean of the simulation results.

Table 1 Proportion of different classes (from Figure 2) of ordered histories for a given $tMRCA_{AB}$

$tMRCA_{AB}$	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
0.1	0.606	0.071	0.284	0.0075	0.015	0.015
0.2	0.366	0.098	0.406	0.027	0.051	0.052
0.4	0.142	0.1	0.415	0.07	0.136	0.137
0.6	0.05	0.083	0.336	0.106	0.211	0.214
0.8	0.02	0.061	0.235	0.141	0.272	0.271
1	0.007	0.04	0.174	0.157	0.301	0.321
1.2	0.002	0.027	0.123	0.17	0.341	0.337
1.4	0	0.02	0.087	0.18	0.359	0.354
1.6	0	0.014	0.053	0.191	0.377	0.365
1.8	0	0.008	0.036	0.19	0.373	0.392

It is interesting to consider what happens when, instead of sampling 2 individuals A and B, we sample k individuals ($k > 2$) to begin with. Imagine that we now know the genealogy of all k individuals, including the times of all coalescent events. As before, we sample 2 new individuals, C and D, and ask what can be said about $tMRCA_{CD}$ in the light of the knowledge of the genealogy of k previously sampled individuals. We know that, as k increases, the probability that the time to the MRCA of the entire population, $tMRCA^*$, equals the $tMRCA$ of the sample is $(k-1)/(k+1)$ (Saunders et al. 1984). For instance, with $k = 19$, the probability that the MRCA of the sample is the same as the MRCA of the population is 0.9, and with $k = 39$, the probability is 0.95. Hence, as k increases, the probability that $tMRCA_{CD}$ will be greater than the $tMRCA$ for all k (previously sampled) individuals will tend to 0. If we only consider 2 most distantly related individuals from amongst the k individuals and call these A and B, then $tMRCA_{AB} \rightarrow tMRCA^*$ as k increases. In this case, $tMRCA_{CD} \leq tMRCA_{AB}$ and we can use the last 3 terms of Equation S16 in the [Supplementary Material online](#) to obtain $E[tMRCA_{CD} | tMRCA_{AB}]$ in the limit as $k \rightarrow \infty$. If the last 3 terms of Equation S16 are used to plot $E[tMRCA_{CD} | tMRCA_{AB}]$ against $tMRCA_{AB} (= tMRCA^*)$, the relationship is very close to a straight line. This, at least, accords with our intuition: the time to common ancestry of any 2 individuals is likely to be related to the time of common ancestry of the whole population.

Up to this point, we have discussed the conditional dependence of times to common ancestry within a population where the genealogies of individuals are described by the coalescent. Nonetheless, entanglement should exert its effect whenever the units we study have an underlying tree-like history and branching events occur stochastically at some well-defined rate that applies across all lineages. Under these conditions, a given time to common ancestry will tend to exert an influence on other times, depending on the genealogy, by constraining the ranges or the limits of these branching times. In this regard, it should not matter whether the units are genes, individuals, species, or higher taxa. We illustrate this by conducting the same simulation as before, except that now the ordered histories of A, B, C, and D are simulated with times to ancestry modeled as a birth–death process, with a combination of different birth and death parameters.

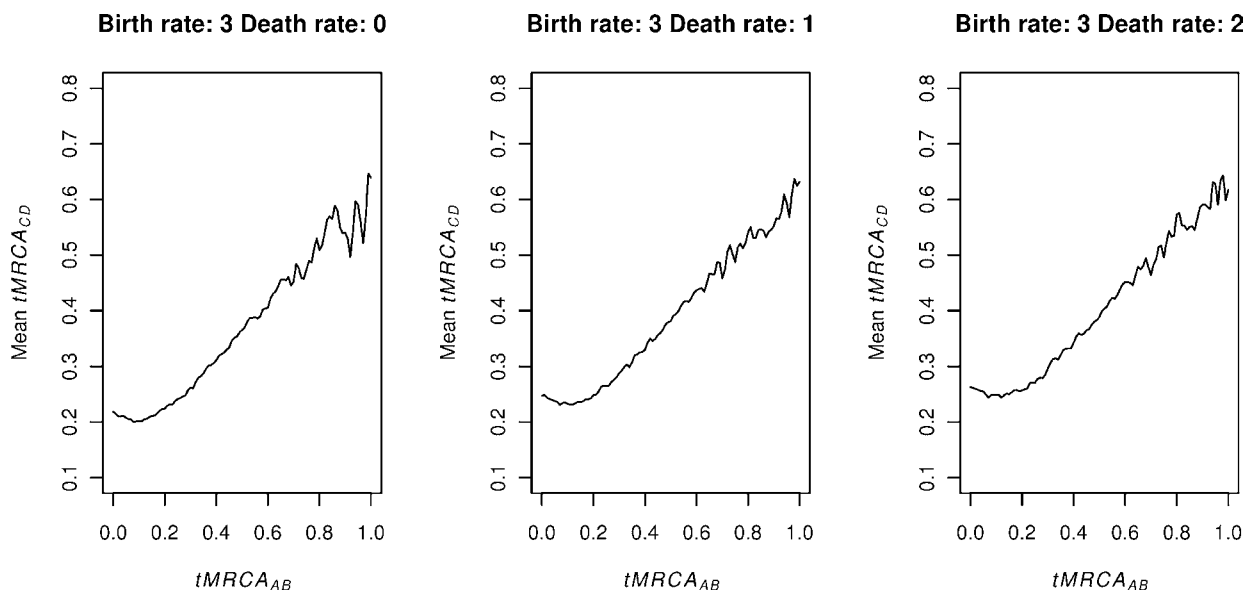


Figure 3. Plot of $E[tMRCA_{CD} | tMRCA_{AB}]$ against $tMRCA_{AB}$ for phylogenies simulated under a birth–death process, with different combinations of birth and death rates. For each combination, 10^6 simulations were performed using R. The axes are scaled in units of time.

Again, it is apparent from the simulations (Figure 3) that $E[tMRCA_{CD} | tMRCA_{AB}]$ depends on $tMRCA_{AB}$.

As noted above, coalescent entanglement is not only a consequence of those ordered histories in which the lineages of A and B intersect those of C and D but also applies when there are no common lineages between (A, B) and (C, D). It is possible, however, that the times to common ancestry of these “phylogenetically independent” pairs are also correlated. To illustrate this effect, we isolated only those histories in Classes 1, 2, and 4 (where pairs [A, B] and [C, D] have no common lineages), and plotted $tMRCA_{CD}$ against $tMRCA_{AB}$ for both coalescent and Yule simulations in which “births” of lineages occur at some nonzero rate, without any lineage “deaths” (Figure 4). Note that in these simulations, the diagonal line (representing genealogies where both [A, B] and [C, D] have the same MRCA) is no longer present. Nonetheless, we again see that for both the coalescent and Yule simulations, $E[tMRCA_{CD} | tMRCA_{AB}]$ varies with $tMRCA_{AB}$.

Discussion

Things that evolve—genes, viruses, organisms, species, languages, and cultures—share common histories. It should come as no surprise that 2 pairs of apparently unrelated units that are part of the same evolving system share information about their entangled ancestries. Indeed, it is precisely for this reason that statistical procedures that take account of underlying evolutionary histories have been developed to analyze the variation in measurable biological quantities, whether these relate to the genetic correlates of disease (Hughes et al. 2008) or the association of different phenotypic traits (Eyre-Walker 2010). It is important to note that entanglement is not only a

consequence of the fact that genealogies contain information about shared evolutionary history. Entanglement is also the result of a functional relationship between each interval of time between coalescent events (=times to common ancestry) and the previous such interval. The relationship between successive coalescent intervals imposes the correlation structure on the times to common ancestry across the genealogy. Recently, Heled and Drummond (2012), in a study on a related issue, explored the mechanics of Bayesian divergence time estimation when the divergence time is known on 1 node of a 3-taxon tree evolving under a Yule process. In their analysis, they show that the prior distribution of times on the node where no divergence time estimate is available, is not independent of the known time of divergence for the other node.

Here, we have only studied the effects of coalescent entanglement in a simple, constant-sized population model. One may speculate on the strength of the entanglement in other populations. For instance, with a growing population, we hypothesize that the correlation between $tMRCA_{CD}$ and $tMRCA_{AB}$ may be stronger, because both (A, B) and (C, D) will tend to have long terminal branches, and coalesce closer to the root (Slatkin and Hudson 1991). When the population is subdivided, the effects of entanglement are likely to depend on the rates of migration between demes and the sizes of the subpopulations from which A, B, C, and D are drawn. In some instances, it is difficult to intuit the effects of entanglement—what would we expect in different selective regimes, for example?

Coalescent entanglement adds another dimension to the extraordinary utility of the coalescent in thinking about the identity and diversity of individuals in a population. In this case, the relationships of distant individuals somehow tell us something about relationships closer to home, for example, the ancestors we share with our spouses and friends. In fact,

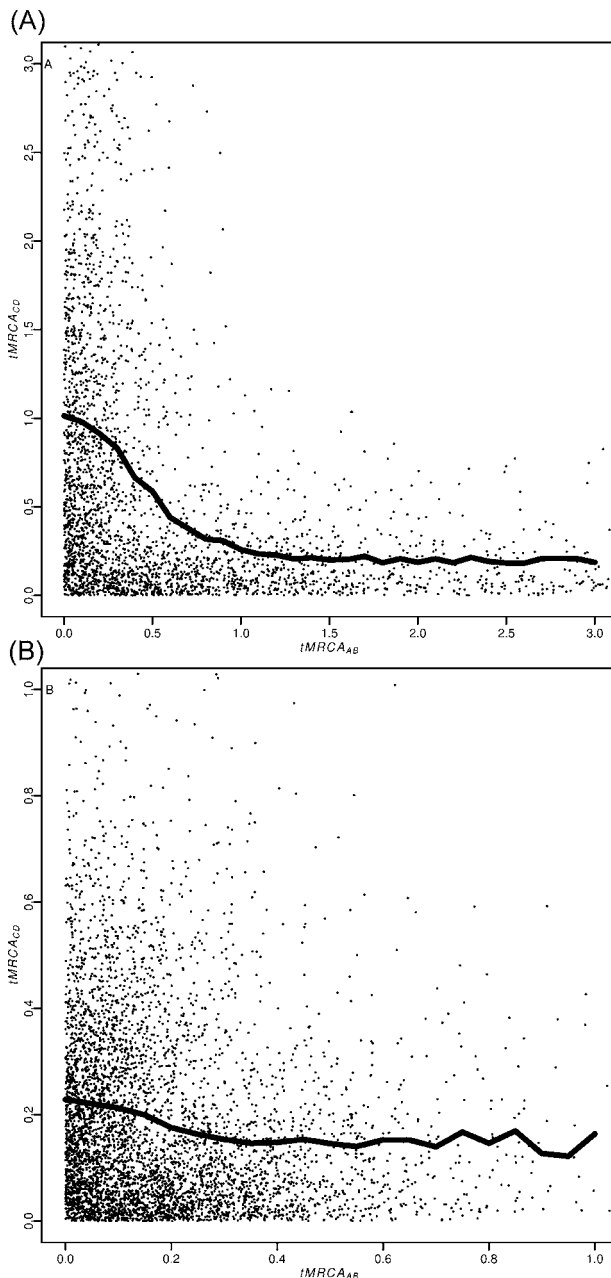


Figure 4. Plot of $tMRCA_{CD}$ against $tMRCA_{AB}$ for ordered histories of Classes 1, 2, and 4 (Figure 2) simulated under (A) the coalescent with axes scaled in N generation units and (B) a birth–death process with birth rate of 3 and death rate of 0 (=a Yule process) with axes scaled in units of time. The bold lines represent $E[tMRCA_{CD} | tMRCA_{AB}]$ for a given $tMRCA_{AB}$ obtained by averaging values of $tMRCA_{CD}$ obtained in the interval $tMRCA_{AB} \pm 0.001$.

the probability that 2 randomly chosen individuals share the same common ancestor as 2 other individuals, sampled previously at random, may be as high as 30–40% (see Table 1, relative frequencies for Class 6 ordered histories). In a letter to Max Born, Einstein, who was not enamored with the idea of quantum entanglement, referred to it as “*spukhafte*

Fernwirkung” or “spooky action at a distance.” If we accept the analogy, we may think of coalescent entanglement as “spooky relations at a distance.”

Supplementary Material

Supplementary material can be found at <http://www.jhered.oxfordjournals.org/>.

Acknowledgments

We thank David Rasmussen, Stacy Scholle, Shishi Luo, Dan Vasco, Joe Felsenstein, and Jotun Hein for helpful comments when we began thinking about this issue. Brian O’Meara helped us understand the consequences of correlated times. The comments of 3 anonymous reviewers and the Associate Editor, Howard Ross, also helped us improve the clarity of the manuscript.

References

- Eyre-Walker A. 2010. Evolution in health and medicine Sackler colloquium: genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proc Natl Acad Sci USA*. 107(Suppl 1):1752–1756.
- Heled J, Drummond AJ. 2012. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Syst Biol*. 61:138–149.
- Hobolth A, Christensen OF, Mailund T, Schierup MH. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet*. 3:e7.
- Hudson RR, Kaplan NL. 1988. The coalescent process in models with selection and recombination. *Genetics*. 120:831–840.
- Hughes AL, Welch R, Puri V, Matthews C, Haque K, Chanock SJ, Yeager M. 2008. Genome-wide SNP typing reveals signatures of population history. *Genomics*. 92:1–8.
- Kaplan NL, Darden T, Hudson RR. 1988. The coalescent process in models with selection. *Genetics*. 120:819–829.
- Kingman JFC. 1982. On the genealogy of large populations. *J Appl Probab*. 19A:27–43.
- Kondo B, Peters JL, Rosensteel BB, Omland KE. 2008. Coalescent analyses of multiple loci support a new route to speciation in birds. *Evolution*. 62:1182–1191.
- McVean GA, Cardin NJ. 2005. Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci*. 360:1387–1393.
- Morlon H, Potts MD, Plotkin JB. 2010. Inferring the dynamics of diversification: a coalescent approach. *PLoS Biol*. 8:e1000493.
- Nath HB, Griffiths RC. 1993. The coalescent in two colonies with symmetric migration. *J Math Biol*. 31:841–851.
- O’Fallon BD, Seger J, Adler FR. 2010. A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies. *Mol Biol Evol*. 27:1162–1172.
- Saunders IW, Tavaré S, Watterson GA. 1984. On the genealogy of nested subsamples from a haploid population. *Adv Appl Probab*. 16:471–491.
- Slatkin M, Hudson RR. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*. 129:555–562.
- Wakeley J, Lessard S. 2006. Corridors for migration between large subdivided populations, and the structured coalescent. *Theor Popul Biol*. 70:412–420.

Received December 14, 2011; Revised July 26, 2012;
Accepted July 31, 2012

Corresponding Editor: Howard Ross