A Coalescent Analysis of Mutation Sharing[1]

We present a coalescent analysis of the distributions of branch- and mutation-sharing under three possible demographic histories. The scenarios increase in complexity. The first is the classical set-up for the 3-coalescent. Here, the sample is exchangeable, representing 3 present-day locii with a selectively neutral history of evolution. In the second, the analysis is extended to allow two locii to have been sampled at (possibly different) earlier times. The third allows differences in sampling time and also relaxes the assumption of panmixia. It builds on the analysis of the first two cases and is the model used in the paper and the analysis implemented in `likelihood_plot.py`.

In all cases we refer to the modern locus as 0 and the potentially ancient locii as 1 and 2 (see Figure 0.2). We are interested in determining the distribution of shared branch lengths between locii 0 and 1, and the probability of a mutation occurring on that branch. We refer to the length variable as L_{01} and the variable number of mutations on the branch as M_{01}. Note that we only consider mutations in the last case, the first two being largely illustrative. By analogy, this analysis also gives us L_{02}. We follow a standard notation in coalescent theory, scaling time in units of $2N_e$ and referring to the random time to first coalescence in the 2- and 3-coalescent processes as T_2 and T_3 respectively [1, 2, 3].

Case 1: Simultaneous sampling and panmixia It is equally likely that any of the three possible pairs of lineages coalesces after some time T_3. Until this event, no ancestral branches are shared and only if the first coalescence is between the lineages of locii 0 and 1 will L_{01} be non-zero. If we condition on this event, we obtain

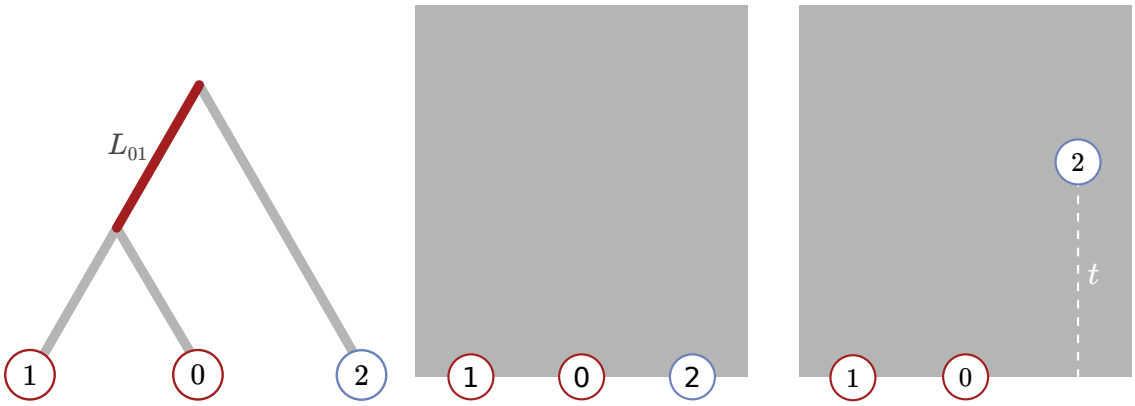
$$\mathbb{P}(L_{01} = l) := f_{01}^1(l) = \begin{cases} \frac{2}{3} & \text{if} \quad l = 0 \\ \frac{1}{3}e^{-l} & \text{if} \quad l > 0. \end{cases} \tag{0.1}$$

The second expression is the chance that the tree topology is as desired multiplied by the probability expression given by the distribution of T_2, the waiting time between the last coalescent events (see Figure 0.1).

Case 2: Staggered sampling and panmixia We look only at the case in which locus 2 is sampled in the past (see Figure 0.1). If we allow both locii 1 and 2 to be ancient sequences, with potentially different sampling times, the analysis is in fact the same: if 1, for example, is more recently sampled, then the problem reduces to the situation described below since no coalescence can occur before another sample enters the population. Observe that this can also be interpreted as a history in which 0 evolves in a panmictic population and the other locii evolve in parallel populations, migrating into the population of 0 at different historical times.

Assume locus 2 is sampled at time t in the past. No ancestral lineage can coalesce with the lineage of 2 before t. We proceed by conditioning on the event that the lineages of 1 and 0 coalesce before then. Let this event be $C_t = \{S < t\}$, where S is a T_2 waiting time. Note that if C_t does not occur (an event we designate

[1] This text is part of a larger work which will be presented in a forthcoming (2018) paper by Tariq Desai and Aylwyn Scally.



(a) Branch to be detected **(b)** Case 1 history **(c)** Case 2 history

Figure 0.1

---

with the symbol $\bar{C}_t$), then the problem reduces to Case 1. Thus

$$\begin{aligned}\mathbb{P}(L_{01} = l) := f^2_{01}(l) &= \mathbb{P}(C_t)\mathbb{P}(L_{01} = l|C_t) + \mathbb{P}(\bar{C}_t)\mathbb{P}(L_{01} = l|\bar{C}_t) \\ &= \left(1 - e^{-t}\right)\mathbb{P}(L_{01} = l|C_t) + e^{-t} f^1_{01}(l).\end{aligned}$$

We evaluate the first term by observing that when $S < t$,

$$L_{01} = K + R \tag{0.2}$$

where $K = t - S^*$ and R is a T_2 waiting time (see Figure 0.2). In this conditional case S^* is distributed according to the truncated exponential $f_{S^*}(s) = e^{-s}/(1 - e^{-t})$. Since K and R are independent, their joint density is

$$f_{K,R}(k, r) = \frac{e^{k-t-r}}{1 - e^{-t}} \qquad 0 < k < t, \quad 0 < r. \tag{0.3}$$

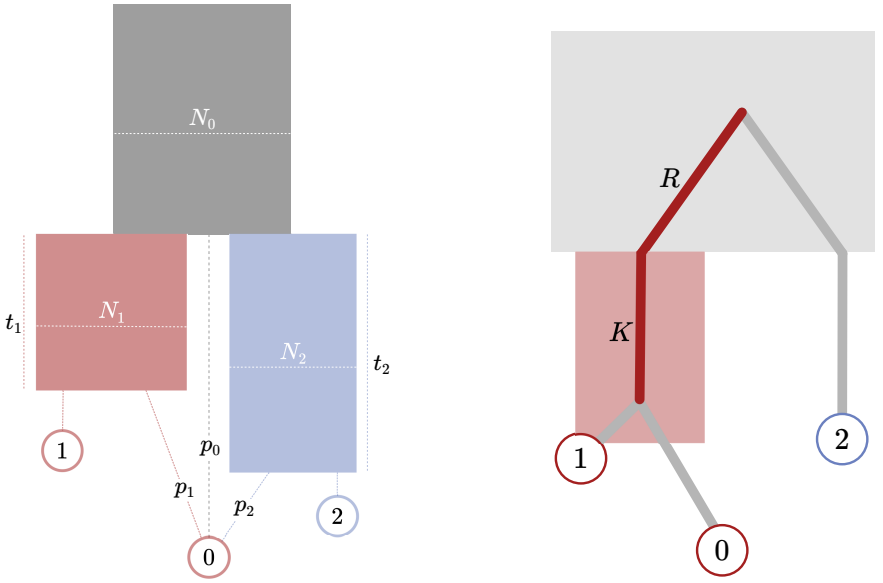
Note that (0.2) implies $K \leq L_{01}$. Coupled with (0.3), this determines the convolution

$$\mathbb{P}(L_{01} = l|C_t) = \int_0^{\min\{l,t\}} f_{K,R}(k, l - k)\, \mathrm{d}k.$$

Evaluating this and combining it with the conditioned expressions, we obtain

$$f^2_{01}(l) = \begin{cases} \frac{2}{3}e^{-t} & \text{if} \quad l = 0 \\ e^{-t-l}\left(\frac{1}{2}\left(e^{2\min\{l,t\}} - 1\right) + \frac{1}{3}\right) & \text{if} \quad l > 0. \end{cases} \tag{0.4}$$

Case 3: Staggered sampling and population substructure We now set up the model in the following way (see Figure 0.2). Allow the ancient sequences 1 and 2 to have derived from populations which exchange



(a) Case 3 history **(b)** Branch to be detected

Figure 0.2

---

no migrants. At some time in the past these populations merge into population 0. Locus 0 enters population 1, 2, or neither, with different probabilities. For some time it might coalesce with the ancient locus in the relevant subpopulation, otherwise it can only do so after the merge time, in which the case the problem reduces to Case 1. Despite the additional parameters, the analysis is very similar to that of Case 2.

We will use the following notation:

$N_i = \lambda_i N_0$	the effective population size of population i
P_i	the event that path i is followed
p_i	the probability of P_i
t_i	the time during which 0 can coalesce with i in population i if P_i occurs
C_i	the event that coalescence occurs between locus i and 0 in population i

The argument again proceeds by conditioning on coalescence occuring within some time window. This time, however, we also need to accomodate population structure, thus additional terms are needed to model the paths lineage 0 might take. First consider the case in which $L_{01} = l$, where $l > 0$:

$$\mathbb{P}(L_{01} = l) = p_1 \mathbb{P}(C_1|P_1)\mathbb{P}(L_{01} = l|C_1) + f_{01}^1(l)\left[p_0 + p_1\mathbb{P}(\bar{C_1}|P_1) + p_2\mathbb{P}(\bar{C_2}|P_2)\right] \tag{0.5}$$

The first term is the probability that locus 0 takes path 1 and coalesces with locus 1 before the merge time. Here $\mathbb{P}(C_1|P_1) = 1 - \exp(-t_1/\lambda_1)$. On the other hand, the second term is the probability that no coalescence occurs before the merge time. We see that $\mathbb{P}(\bar{C_1}|P_1) = \exp(-t_1/\lambda_1)$ and similarly, $\mathbb{P}(\bar{C_2}|P_2) = \exp(-t_2/\lambda_2)$.

All that remains to calculate is $\mathbb{P}(L_{01} = l|C_1)$. The argument is as before. Let $L_{01} = K + R$ where R is a

T_2 waiting time and $K = t_1 - S^*$ where S^* is distributed according to a truncated exponential. Adjusting for the difference in population size, the joint distribution of independent variables K and R is

$$f_{K,R}(k,r) = \frac{\lambda_1^{-1} e^{(k-t_1)/\lambda_1 - r}}{1 - e^{t_1/\lambda_1}} \qquad 0 < k < t_1, \quad 0 < r.$$

The convolution, as before, is readily determined. Combining it with expression (0.5), we obtain

$$\begin{aligned} \mathbb{P}(L_{01} = l) &= A_1 e^{-l} \left(e^{(1+1/\lambda_1)\min\{t_1, l\}} - 1 \right) + (1/3) B e^{-l} \\ \text{where} \quad A_1 &= \frac{p_1 e^{-t_1/\lambda_1}}{\lambda_1 + 1}, \\ \text{and} \quad B &= p_0 + p_1 e^{-t_1/\lambda_1} + p_2 e^{-t_2/\lambda_2}. \end{aligned} \tag{0.6}$$

Another exercise in branch-accounting provides the probability

$$\begin{aligned} \mathbb{P}(L_{01} = 0) &= p_2 \mathbb{P}(C_2 | P_2) + f_{01}^1(0) \left[p_0 + p_1 \mathbb{P}(\bar{C}_1 | P_1) + p_2 \mathbb{P}(\bar{C}_2 | P_2) \right] \\ &= p_2 (1 - e^{-t_2/\lambda_2}) + (2/3) B. \end{aligned} \tag{0.7}$$

For the purposes of testing, we can express these probabilities as the cumulative distribution function

$$F_{01}^3(l) = \begin{cases} 0 & \text{if} \quad l < 0 \\ p_2 (1 - e^{-t_2/\lambda_2}) + (2/3) B & \text{if} \quad l = 0 \\ F_{01}^3(0) + A_1 \left(\lambda_1 e^{l/\lambda_1} + e^{-l} - \lambda_1 - 1 \right) + (1/3) \left(1 - e^{-l} \right) B & \text{if} \quad 0 < l < t_1 \\ F_{01}^3(t_1) + \left(A_1 \left(e^{(1+1/\lambda_1) t_1} - 1 \right) + (1/3) B \right) \left(e^{-t_1} - e^{-l} \right) & \text{if} \quad t_1 < l. \end{cases} \tag{0.8}$$

Detecting segment matches If it were possible to extract marginal genealogical trees from an accurately inferred ARG, we might be able to infer demography based purely on the joint distribution of shared branches. Since these branches are not directly observed, we are interested in the probability that locii share "private" derived alleles. In the case we have been working with so far, this is the same as asking whether locii 0 and 1 share any mutations which are not shared with 2. Under the infinite-sites assumption this is possible if and only if the lineages of 0 and 1 coalesce first. If this kind of mutation has occured, we say that a segment match has been made, or that some branch-sharing is *detectable*.

Under the models we have considered thus far, the probability of detecting branch-sharing can be analytically determined. We demonstrate this in the third case. If M_{01} is the random number of shared private mutations of locii 0 and 1, then a segment matches if $M_{01} > 0$, and thus the relevant quantity to determine is $1 - \mathbb{P}(M_{01} = 0)$. Expressions (0.7) and (0.6) allow us to get this by conditioning on shared branch-lengths. In other words,

$$\mathbb{P}(M_{01} = 0) = 1.\mathbb{P}(L_{01} = 0) + \int_0^\infty \mathbb{P}(M_{01} = 0 | L_{01} = l) \mathbb{P}(L_{01} = l) \, \mathrm{d}l. \tag{0.9}$$

(With a slight abuse of notation we assume, in the second term, that $l > 0$ to avoid including the point mass at $l = 0$. We also suppress the dependence on demographic parameters.) We use the standard coalescent model

of mutation and assume that the random number of mutations on some branch is governed by a Poisson process with intensity parameter $\theta/2$. In our case $\theta = 4N_0\mu b$, with μ being the per site per generation mutation rate and b being the number of sites in our locus. The resulting integral is straightforward to evaluate, though is perhaps easiest to express as a sum of three terms:

$$\begin{aligned} I &= \int_0^\infty e^{-l\theta/2}\left[A_1 e^{-l}\left(e^{(1+1/\lambda_1)\min\{t_1,l\}}-1\right)+B(1/3)e^{-l}\right]\mathrm{d}l \\ &= I_1 + I_2 + I_3 \end{aligned}$$

$$\begin{aligned} \text{where} \quad I_1 &= A_1\left[\frac{e^{(1/\lambda_1-\theta/2)t_1}-1}{1/\lambda_1-\theta/2}+\frac{e^{-(1+\theta/2)t_1}-1}{1+\theta/2}\right], \\ I_2 &= A_1\left(e^{(1+1/\lambda_1)t_1}-1\right)\left(\frac{e^{-(1+\theta/2)t_1}}{1+\theta/2}\right), \\ \text{and} \quad I_3 &= \frac{B}{3(1+\theta/2)}. \end{aligned} \tag{0.10}$$

Using this notation, the probability that we detect a match between locii 0 and 1 under the third demographic scenario is given by the expression:

$$\mathbb{P}(M_{01}>0) = 1-\left(F_{01}^3(0)+I_1+I_2+I_3\right). \tag{0.11}$$

The relevant probability for the likelihood analysis is this value as a proportion of the segments expected to match with either locus 1 or 2, ie.

$$\frac{\mathbb{P}(M_{01}>0)}{\mathbb{P}(M_{01}>0)+\mathbb{P}(M_{02}>0)}. \tag{0.12}$$

Several extensions of these analyses are obvious next steps. For instance, you might estimate the distributions of any specific number of shared mutations. More complicated demographic models might also be considered.

References

[1] J. F. C. Kingman, "On the Genealogy of Large Populations Author(s): J. F. C. Kingman Source: Journal of Applied Probability, Vol. 19, Essays in Statistical Science (1982), pp. 27-43 Published by:," *Journal of Applied Probability*, vol. 19, pp. 27–43, 1982.

[2] J. Hein, M. Schierup, and C. Wiuf, *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory.* Oxford University Press, USA, 2004.

[3] J. Wakeley, *Coalescent Theory: An Introduction.* Macmillan Learning, 2009.