

SUPPORTING INFORMATION

GENETIC VARIABILITY UNDER THE SEEDBANK COALESCENT

J. BLATH, A. GONZÁLEZ CASANOVA, B. ELDON, N. KURT,
M. WILKE BERENGUER

File S1 Proofs and further recursive formulas

Expectation and variance of the T_{MRCA}

For $n, m \in \mathbb{N}_0$, let $t_{n,m} := \mathbb{E}_{n,m}[T_{\text{MRCA}}]$ and $v_{n,m} := \mathbb{V}_{n,m}[T_{\text{MRCA}}]$.

Proposition S1.1. *Let $n, m \in \mathbb{N}_0$. Then we have the following recursive representations*

$$\mathbb{E}_{n,m}[T_{\text{MRCA}}] = t_{n,m} = \lambda_{n,m}^{-1} + \alpha_{n,m}t_{n-1,m} + \beta_{n,m}t_{n-1,m+1} + \gamma_{n,m}t_{n+1,m-1}, \quad (\text{S1})$$

$$\begin{aligned} \mathbb{V}_{n,m}[T_{\text{MRCA}}] = v_{n,m} &= \lambda_{n,m}^{-2} + \alpha_{n,m}v_{n-1,m} + \beta_{n,m}v_{n-1,m+1} + \gamma_{n,m}v_{n+1,m-1} \\ &\quad + \alpha_{n,m}t_{n-1,m}^2 + \beta_{n,m}t_{n-1,m+1}^2 + \gamma_{n,m}t_{n+1,m-1}^2 \\ &\quad - \left(\alpha_{n,m}t_{n-1,m} + \beta_{n,m}t_{n-1,m+1} + \gamma_{n,m}t_{n+1,m-1}\right)^2, \end{aligned} \quad (\text{S2})$$

with initial conditions $t_{1,0} = t_{0,1} = v_{1,0} = v_{0,1} = 0$.

Proof of Proposition S1.1. Let τ_1 denote the time of the first jump of the process $(N_t, M_t)_{t \geq 0}$. If started at (n, m) , this is an exponential random variable with parameter $\lambda_{n,m}$. Applying the strong Markov property we obtain

$$\begin{aligned} t_{n,m} &= \mathbb{E}_{n,m}[\tau_1] + \mathbb{E}_{n,m}[\mathbb{E}_{N_{\tau_1}, M_{\tau_1}}[T_{\text{MRCA}}]] \\ &= \lambda_{n,m}^{-1} + \alpha_{n,m}t_{n-1,m} + \beta_{n,m}t_{n-1,m+1} + \gamma_{n,m}t_{n+1,m-1}. \end{aligned}$$

Similarly, the strong Markov property (telling us that τ_1 is independent of the time to the most recent common ancestor of the (random) sample (N_{τ_1}, M_{τ_1})) and the law of total

variance yields

$$\begin{aligned} v_{n,m} &= \mathbb{V}_{n,m}[\tau_1] + \mathbb{E}_{n,m}[\mathbb{V}_{N_{\tau_1}, M_{\tau_1}}[T_{\text{MRCA}}]] + \mathbb{V}_{n,m}[\mathbb{E}_{N_{\tau_1}, M_{\tau_1}}[T_{\text{MRCA}}]] \\ &= \lambda_{n,m}^{-2} + \mathbb{E}_{n,m}[\mathbb{V}_{N_{\tau_1}, M_{\tau_1}}[T_{\text{MRCA}}]] + \mathbb{V}_{n,m}[\mathbb{E}_{N_{\tau_1}, M_{\tau_1}}[T_{\text{MRCA}}]]. \end{aligned}$$

We have

$$\mathbb{E}_{n,m}[\mathbb{V}_{N_{\tau_1}, M_{\tau_1}}[T_{\text{MRCA}}]] = \alpha_{n,m}v_{n-1,m} + \beta_{n,m}v_{n-1,m+1} + \gamma_{n,m}v_{n+1,m-1}$$

and

$$\begin{aligned} \mathbb{V}_{n,m}[\mathbb{E}_{N_{\tau_1}, M_{\tau_1}}[T_{\text{MRCA}}]] &= \mathbb{E}_{n,m}[\mathbb{E}_{N_{\tau_1}, M_{\tau_1}}[T_{\text{MRCA}}]^2] - \mathbb{E}_{n,m}[\mathbb{E}_{N_{\tau_1}, M_{\tau_1}}[T_{\text{MRCA}}]]^2 \\ &= \alpha_{n,m}t_{n-1,m}^2 + \beta_{n,m}t_{n-1,m+1}^2 + \gamma_{n,m}t_{n+1,m-1}^2 \\ &\quad - (\alpha_{n,m}t_{n-1,m} + \beta_{n,m}t_{n-1,m+1} + \gamma_{n,m}t_{n+1,m-1})^2. \end{aligned}$$

Combining the observations proves the result. \square

Expectation and variance of the total tree length

Let $l_{n,m}^{(a)} := \mathbb{E}_{n,m}[L^{(a)}]$ and $l_{n,m}^{(d)} := \mathbb{E}_{n,m}[L^{(d)}]$ denote the expectations, and $w_{n,m}^{(a)} := \mathbb{V}_{n,m}[L^{(a)}]$ and $w_{n,m}^{(d)} := \mathbb{V}_{n,m}[L^{(d)}]$ the variances of the total tree lengths, and define the mixed second moment, $w_{n,m}^{(a,d)} := \mathbb{E}_{n,m}[L^{(a)}L^{(d)}]$.

Proposition S1.2 (Recursion: Total tree length). *For $n, m \in \mathbb{N}$ we have*

$$l_{n,m}^{(a)} = n\lambda_{n,m}^{-1} + \alpha_{n,m}l_{n-1,m}^{(a)} + \beta_{n,m}l_{n-1,m+1}^{(a)} + \gamma_{n,m}l_{n+1,m-1}^{(a)} \quad (\text{S3})$$

$$l_{n,m}^{(d)} = m\lambda_{n,m}^{-1} + \alpha_{n,m}l_{n-1,m}^{(d)} + \beta_{n,m}l_{n-1,m+1}^{(d)} + \gamma_{n,m}l_{n+1,m-1}^{(d)}, \quad (\text{S4})$$

and

$$\begin{aligned}
w_{n,m}^{(a)} &= n^2 \lambda_{n,m}^{-2} + \alpha_{n,m} w_{n-1,m}^{(a)} + \beta_{n,m} w_{n-1,m+1}^{(a)} + \gamma_{n,m} w_{n+1,m-1}^{(a)} \\
&\quad + \alpha_{n,m} (l_{n-1,m}^{(a)})^2 + \beta_{n,m} (l_{n-1,m+1}^{(a)})^2 + \gamma_{n,m} (l_{n+1,m-1}^{(a)})^2 \\
&\quad - \left(\alpha_{n,m} l_{n-1,m}^{(a)} + \beta_{n,m} l_{n-1,m+1}^{(a)} + \gamma_{n,m} l_{n+1,m-1}^{(a)} \right)^2,
\end{aligned} \tag{S5}$$

$$\begin{aligned}
w_{n,m}^{(d)} &= m^2 \lambda_{n,m}^{-2} + \alpha_{n,m} w_{n-1,m}^{(d)} + \beta_{n,m} w_{n-1,m+1}^{(d)} + \gamma_{n,m} w_{n+1,m-1}^{(d)} \\
&\quad + \alpha_{n,m} (l_{n-1,m}^{(d)})^2 + \beta_{n,m} (l_{n-1,m+1}^{(d)})^2 + \gamma_{n,m} (l_{n+1,m-1}^{(d)})^2 \\
&\quad - \left(\alpha_{n,m} l_{n-1,m}^{(d)} + \beta_{n,m} l_{n-1,m+1}^{(d)} + \gamma_{n,m} l_{n+1,m-1}^{(d)} \right)^2,
\end{aligned} \tag{S6}$$

$$w_{n,m}^{(a,d)} = 2nm \lambda_{n,m}^{-2} + \alpha_{n,m} w_{n-1,m}^{(a,d)} + \beta_{n,m} w_{n-1,m+1}^{(a,d)} + \gamma_{n,m} w_{n+1,m-1}^{(a,d)}. \tag{S7}$$

Proof of Proposition S1.2. The result can easily be obtained observing that each stretch of time of length τ in which we have a constant number of n active blocks and m dormant blocks contributes with $n\tau$ to the total active tree length, and with $m\tau$ to the total dormant tree length. Thus we have

$$l_{n,m}^{(a)} = n \mathbb{E}_{n,m}[\tau_1] + \mathbb{E}_{n,m}[\mathbb{E}_{N_{\tau_1}, M_{\tau_1}}[L^{(a)}]],$$

and we proceed as in the proof of Proposition S1.1. From these quantities we easily obtain the expected total tree length as $l_{n,m}^{(a)} + l_{n,m}^{(d)}$. Moreover,

$$\text{Cov}_{n,m}(L^{(a)}, L^{(d)}) = w_{n,m}^{(a,d)} - w_{n,m}^{(a)} w_{n,m}^{(d)}.$$

□

Expectation of total length of external branches

To derive recursions for the total length of external branches in either of the two states is a little more involved, since obviously a coalescence can happen between either two external active branches, two internal active branches, or an external and an internal active branch. We use indices (n, n', m, m') to denote the number of external active branches, internal active branches, external dormant branches, and internal dormant branches, respectively. Abbreviate

$$\begin{aligned}\alpha_{n,n',m,m'}^{(1)} &:= \frac{\binom{n}{2}}{\lambda_{n+n',m+m'}}, & \alpha_{n,n',m,m'}^{(2)} &:= \frac{\binom{n'}{2}}{\lambda_{n+n',m+m'}}, & \alpha_{n,n',m,m'}^{(3)} &:= \frac{nn'}{\lambda_{n+n',m+m'}}, \\ \beta_{n,n',m,m'}^{(1)} &:= \frac{cn}{\lambda_{n+n',m+m'}}, & \beta_{n,n',m,m'}^{(2)} &:= \frac{cn'}{\lambda_{n+n',m+m'}}, \\ \gamma_{n,n',m,m'}^{(1)} &:= \frac{cKm}{\lambda_{n+n',m+m'}}, & \gamma_{n,n',m,m'}^{(2)} &:= \frac{cKm'}{\lambda_{n+n',m+m'}}.\end{aligned}$$

Let $E^{(a)}$ denote the total length of external branches in the plant state, and $E^{(d)}$ the total length of external branches in the seed state. Then we have

Proposition S1.3 (Recursion: Total length of external branches). *For $n, m \in \mathbb{N}$, we have the representation*

$$\mathbb{E}_{n,m}[E^{(a)}] = e_{n,0,m,0}^{(a)}, \quad \mathbb{E}_{n,m}[E^{(d)}] = e_{n,0,m,0}^{(d)},$$

where $e_{n,n',m,m'}^{(a)}$ and $e_{n,n',m,m'}^{(d)}$, $n, n', m, m' \in \mathbb{N}_0$ satisfy the recursions

$$\begin{aligned}e_{n,n',m,m'}^{(a)} &= n\lambda_{n+n',m+m'}^{-1} \\ &+ \alpha_{n,n',m,m'}^{(1)}e_{n-2,n'+1,m,m'}^{(a)} + \alpha_{n,n',m,m'}^{(2)}e_{n,n'-1,m,m'}^{(a)} + \alpha_{n,n',m,m'}^{(3)}e_{n-1,n',m,m'}^{(a)} \\ &+ \beta_{n,n',m,m'}^{(1)}e_{n-1,n',m+1,m'}^{(a)} + \beta_{n,n',m,m'}^{(2)}e_{n,n'-1,m,m'+1}^{(a)} \\ &+ \gamma_{n,n',m,m'}^{(1)}e_{n+1,n',m-1,m'}^{(a)} + \gamma_{n,n',m,m'}^{(2)}e_{n,n'+1,m,m'-1}^{(a)}\end{aligned}$$

and

$$\begin{aligned}
e_{n,n',m,m'}^{(d)} &= m\lambda_{n+n',m+m'}^{-1} \\
&+ \alpha_{n,n',m,m'}^{(1)} e_{n-2,n'+1,m,m'}^{(d)} + \alpha_{n,n',m,m'}^{(2)} e_{n,n'-1,m,m'}^{(d)} + \alpha_{n,n',m,m'}^{(3)} e_{n-1,n',m,m'}^{(d)} \\
&+ \beta_{n,n',m,m'}^{(1)} e_{n-1,n',m+1,m'}^{(d)} + \beta_{n,n',m,m'}^{(2)} e_{n,n'-1,m,m'+1}^{(d)} \\
&+ \gamma_{n,n',m,m'}^{(1)} e_{n+1,n',m-1,m'}^{(d)} + \gamma_{n,n',m,m'}^{(2)} e_{n,n'+1,m,m'-1}^{(d)}
\end{aligned}$$

Observing that $e_{0,n',0,m'}^{(a)} = e_{0,n',0,m'}^{(d)} = 0$ for all n', m' , and $e_{1,0,0,0}^{(a)} = e_{1,0,0,0}^{(d)} = 0$, and that the total number $n + n' + m + m'$ is non-increasing, these recursions can be solved iteratively.

Proof of Proposition S1.3. This follows by a similar first-step analysis as in Proposition S1.2, taking into account the transitions for internal and external branches, and observing that at each coalescence event between two external branches, the number of external plant branches is reduced by two and the number of internal branches is increased by one, in a coalescence of an external and an internal branch, the number of external plant branches is reduced by one and the number of internal plant branches stays the same, and in a coalescence of two internal branches, their number is reduced by one. \square

Obviously, the expected total length of external branches is then given by $e_{n,0,m,0}^{(a)} + e_{n,0,m,0}^{(d)}$. Note that proceeding as in Proposition S1.2, we could also give recursions for the variances of these quantities.

Expectation and variance of the number of segregating sites

Proposition S1.4. *For $n, m \in \mathbb{N}_0$ we have*

$$\mathbb{E}_{n,m}[S] = \frac{\theta_1}{2} l_{n,m}^{(a)} + \frac{\theta_2}{2} l_{n,m}^{(d)},$$

and

$$\mathbb{V}_{n,m}[S] = \frac{\theta_1}{2} l_{n,m}^{(a)} + \frac{\theta_2}{2} l_{n,m}^{(d)} + \frac{\theta_1^2}{4} w_{n,m}^{(a)} + \frac{\theta_2^2}{4} w_{n,m}^{(d)} + \frac{\theta_1 \theta_2}{2} (w_{n,m}^{a,d} - l_{n,m}^{(a)} l_{n,m}^{(d)}),$$

where $l_{n,m}^{(a)}, l_{n,m}^{(d)}, w_{n,m}^{(a)}, w_{n,m}^{(d)}$ and $w_{n,m}^{(a,d)}$ are given by Proposition S1.2.

Proof of Proposition S1.4. Observe that conditional on the total lengths $L^{(a)}, L^{(d)}$, the number of segregating sites is the sum of two independent Poisson random variables with parameters $\theta_1 L^{(a)}/2$ and $\theta_2 L^{(d)}/2$, respectively. Hence, if an ancestral line is in the plant state for a period of time of length $L > 0$, the expected number of mutations that occur in this period is $L\theta_1/2$. Similarly, in a period of length L when the ancestral line is a seed, the expected number of mutations is $L\theta_2/2$. Thus the first result follows directly from Proposition S1.2.

For the second result, we apply the law of total variance and obtain similarly that

$$\begin{aligned} \mathbb{V}_{n,m}(S) &= \mathbb{E}_{n,m}[\mathbb{V}(S \mid L^{(a)}, L^{(d)})] + \mathbb{V}_{n,m}(\mathbb{E}[S \mid L^{(a)}, L^{(d)}]) \\ &= \mathbb{E}_{n,m} \left[\frac{\theta_1}{2} L^{(a)} + \frac{\theta_2}{2} L^{(d)} \right] + \mathbb{V}_{n,m} \left(\frac{\theta_1}{2} L^{(a)} + \frac{\theta_2}{2} L^{(d)} \right) \\ &= \frac{\theta_1}{2} l_{n,m}^{(a)} + \frac{\theta_2}{2} l_{n,m}^{(d)} + \frac{\theta_1^2}{4} w_{n,m}^{(a)} + \frac{\theta_2^2}{4} w_{n,m}^{(d)} + 2 \frac{\theta_1}{2} \frac{\theta_2}{2} \text{Cov}_{n,m}(L^{(a)}, L^{(d)}). \end{aligned}$$

□

It is possible to directly derive a recursion for the number of segregating sites without explicitly passing through calculating the tree lengths. Since it may be of use we state it here. Let

$$s_{n,m} := \mathbb{E}_{n,m}[S], \quad \text{and} \quad z_{n,m} := \mathbb{V}_{n,m}(S).$$

Proposition S1.5 (Alternative recursion). *Let $n, m \in \mathbb{N}_0$. Then*

$$s_{n,m} = \left(\frac{\theta_1}{2} n + \frac{\theta_2}{2} m \right) \lambda_{n,m}^{-1} + \alpha_{n,m} s_{n-1,m} + \beta_{n,m} s_{n-1,m+1} + \gamma_{n,m} s_{n+1,m-1} \quad (\text{S8})$$

$$\begin{aligned} z_{n,m} &= \left(\frac{\theta_1}{2} n + \frac{\theta_2}{2} m \right) \lambda_{n,m}^{-1} + \left(\frac{\theta_1}{2} n + \frac{\theta_2}{2} m \right)^2 \lambda_{n,m}^{-2} \\ &\quad + \alpha_{n,m} z_{n-1,m} + \beta_{n,m} z_{n-1,m+1} + \gamma_{n,m} z_{n+1,m-1} \\ &\quad + \alpha_{n,m} s_{n-1,m}^2 + \beta_{n,m} s_{n-1,m+1}^2 + \gamma_{n,m} s_{n+1,m-1}^2 \\ &\quad - \left(\alpha_{n,m} s_{n-1,m} + \beta_{n,m} s_{n-1,m+1} + \gamma_{n,m} s_{n+1,m-1} \right)^2. \end{aligned} \quad (\text{S9})$$

Proof of Proposition S1.5. Let σ_1 denote the number of mutations that occur until time τ_1 , which was defined in the proof of Proposition S1.1. Given $\tau_1 = t$, we know that σ_1 is the sum of two independent Poisson random variables with parameters $\theta_1 nt$ and $\theta_2 mt$, respectively. As in the previous proof we obtain

$$\begin{aligned} s_{n,m} &= \mathbb{E}_{n,m}[\sigma_1] + \mathbb{E}_{n,m}[\mathbb{E}_{N_{\tau_1}, M_{\tau_1}}[S]] \\ &= \left(\frac{\theta_1}{2}n + \frac{\theta_2}{2}m\right) \mathbb{E}_{n,m}[\tau_1] + \alpha_{n,m}s_{n-1,m} + \beta_{n,m}s_{n-1,m+1} + \gamma_{n,m}s_{n+1,m-1} \end{aligned}$$

and

$$z_{n,m} = \mathbb{V}_{n,m}(\sigma_1) + \mathbb{E}_{n,m}[\mathbb{V}_{N_{\tau_1}, M_{\tau_1}}(S)] + \mathbb{V}_{n,m}(\mathbb{E}_{N_{\tau_1}, M_{\tau_1}}[S]).$$

Once more using the law of total variance we obtain

$$\begin{aligned} \mathbb{V}_{n,m}[\sigma_1] &= \mathbb{E}_{n,m}[\mathbb{V}_{n,m}(\sigma_1 \mid \tau_1)] + \mathbb{V}_{n,m}(\mathbb{E}_{n,m}[\sigma_1 \mid \tau_1]) \\ &= \left(\frac{\theta_1}{2}n + \frac{\theta_2}{2}m\right) \mathbb{E}_{n,m}[\tau_1] + \left(\frac{\theta_1}{2}n + \frac{\theta_2}{2}m\right)^2 \mathbb{V}_{n,m}[\tau_1] \\ &= \left(\frac{\theta_1}{2}n + \frac{\theta_2}{2}m\right) \lambda_{n,m}^{-1} + \left(\frac{\theta_1}{2}n + \frac{\theta_2}{2}m\right)^2 \lambda_{n,m}^{-2}. \end{aligned} \tag{S10}$$

The same calculations as in the proof of Proposition S1.1 lead to

$$\mathbb{E}_{n,m}[\mathbb{V}_{N_{\tau_1}, M_{\tau_1}}(S)] = \alpha_{n,m}z_{n-1,m} + \beta_{n,m}z_{n-1,m+1} + \gamma_{n,m}z_{n+1,m-1},$$

and

$$\begin{aligned} \mathbb{V}_{n,m}(\mathbb{E}_{N_{\tau_1}, M_{\tau_1}}[S]) &= \alpha_{n,m}s_{n-1,m}^2 + \beta_{n,m}s_{n-1,m+1}^2 + \gamma_{n,m}s_{n+1,m-1}^2 \\ &\quad - (\alpha_{n,m}s_{n-1,m} + \beta_{n,m}s_{n-1,m+1} + \gamma_{n,m}s_{n+1,m-1})^2. \end{aligned}$$

□

Expected value of average pairwise differences (π)

Recall the definition (19) of average pairwise difference $\pi = \binom{N_0+M_0}{2}^{-1} \sum_{(i,j): i < j} K_{i,j}$.

Proposition S1.6. *For $n, m \in \mathbb{N}_0$ we have*

$$\mathbb{E}_{n,m}[\pi] = \frac{1}{\binom{n+m}{2}} \left\{ \binom{n}{2} \left(\frac{\theta_1}{2} l_{2,0}^{(a)} + \frac{\theta_2}{2} l_{2,0}^{(d)} \right) + nm \left(\frac{\theta_1}{2} l_{1,1}^{(a)} + \frac{\theta_2}{2} l_{1,1}^{(d)} \right) + \binom{m}{2} \left(\frac{\theta_1}{2} l_{0,1}^{(a)} + \frac{\theta_2}{2} l_{0,1}^{(d)} \right) \right\}$$

Proof of Proposition S1.6. By definition

$$\mathbb{E}_{n,m}[\pi] = \frac{1}{\binom{n+m}{2}} \sum_{1 \leq i < j \leq n+m} \mathbb{E}_{n,m}[K_{i,j}].$$

When comparing two individuals their *pairwise* differences in the infinite sites model coincide with the number of mutations that occurred along the branches of their corresponding sub-tree and are thus given the product of the mutation rate and length of the branches. Therefore, $\mathbb{E}_{n,m}[K_{i,j}]$ actually only depends on whether i, j are dormant or active individuals. We obtain

$$\mathbb{E}_{n,m}[K_{i,j}] = \begin{cases} \frac{\theta_1}{2} l_{2,0}^{(a)} + \frac{\theta_2}{2} l_{2,0}^{(d)}, & \text{if } i, j \text{ are active} \\ \frac{\theta_1}{2} l_{1,1}^{(a)} + \frac{\theta_2}{2} l_{1,1}^{(d)}, & \text{if } i \text{ is active and } j \text{ dormant} \\ \frac{\theta_1}{2} l_{0,2}^{(a)} + \frac{\theta_2}{2} l_{0,2}^{(d)}, & \text{if } i, j \text{ are dormant.} \end{cases}$$

Substituting this into the above equation, the result follows. □