Genealogical distances under low levels of selection

Elisabeth Huss, Peter Pfaffelhuber

December 13, 2018

Abstract

For a panmictic population of constant size evolving under neutrality, Kingman's coalescent describes the genealogy of a population sample in equilibrium. However, for genealogical trees under selection, not even expectations for most basic quantities like height and length of the resulting random tree are known. Here, we give an analytic expression for the distribution of the total tree length of a sample of size n under low levels of selection in a two-alleles model. We can prove that trees are shorter than under neutrality under genic selection and if the beneficial mutant has dominance $h < 1/2$, but longer for $h > 1/2$. The difference to neutrality is $O(\alpha^2)$ for genic selection with selection intensity α and $O(\alpha)$ for other modes of dominance.

1 Introduction

Understanding population genetic models, e.g. the Wright-Fisher or the Moran model, can be achieved in various ways. Classically, allelic frequencies are described by diffusions in the large population limit, and for simple models such as two-alleles models, the theory of one-dimensional diffusions leads to prediction for virtually all quantities of interest (Ewens, 2004). Moreover, starting with Kingman (1982) and Hudson (1983), genealogical trees started to play a big role in the understanding of the models as well as of DNA data from a population sample. Most importantly, all variation seen in data can be mapped onto a genealogical tree. Under neutral evolution, the mutational process is independent of the genealogical tree. As a consequence, the length of the genealogical tree is proportional to the total amount of polymorphic sites in the sample.

Genealogies under selection have long been an interesting object to study (see e.g. Wakeley, 2010 for a review). Starting with Krone and Neuhauser (1997) and Neuhauser and Krone (1997), genealogical trees under selection could be described using the Ancestral Selection Graph (ASG). In addition to coalescence events, the fitness differences make it necessary to study the history of all possible ancestors, leading to splitting events in the ASG. The disadvantage of these splitting events is that they make this genealogical structure far more complicated to study than the coalescent for neutral evolution.

AMS 2010 subject classification. 92D15 (Primary) 60J70 (Secondary).
Keywords and phrases. Coalescent, tree length, dominance coefficient, two-allele-model

In recent years, much progress has been made in the simulation of genealogical trees under selection. Mostly, these simulation algorithms use the approach of the structured coalescent, which is based on Kaplan et al. (1988). Here, the allelic frequency path is generated first, and conditional on this path, coalescence events are carried out. (See also Barton et al., 2004 for a formal derivation of this approach.) Simulation approaches based on this idea include the inference method by Coop and Griffiths (2004), msms by Ewing and Hermisson (2010), and discoal Kern and Schrider (2016). However, the structured coalescent approach can hardly be used to obtain analytical insights (with some notable exceptions, see e.g. Taylor, 2007).

Recently, genealogies under selection have been studied by Depperschmidt et al. (2012) using Markov processes taking values in the space of trees, i.e. the genealogical tree is modelled as a stochastic process which is changing as the population evolves. As for many Markov processes, the equilibrium of this process gives the equilibrium tree and can be studied using stationary solutions of differential equations. In our manuscript, we will make use of this theory in order to compute an approximation for the total tree length under a general bi-allelic selection scheme, which is assumed to be weak; see Section 4. Our results are extensions of Theorem 5 of Depperschmidt et al. (2012), where an approximation of the Laplace-transform of the genealogical distance of a pair of individuals under bi-allelic mutation and low levels of selection was computed.

The paper is structured as follows: In Section 2, we introduce the model we are going to study, i.e. genealogies in the large population limit for a Moran model under genic or incomplete dominance, over- or under-dominant selection. The last three cases we collect under the term *other modes of dominance*. We give recursions for the Laplace transform (and the expectation) of the tree length of a sample in Theorem 1 and Corollary 4 for genic selection, and in Theorem 2 and Corollary 7 for other modes of dominance. In Section 3, we discuss our findings and also provide some plots on the change of tree lengths under selection. Section 4 gives some preliminaries for the proofs. In particular, we give a brief review of the construction of evolving genealogies from Depperschmidt et al. (2012). Finally, Section 5 contains all proofs.

2 Model and main results

We will obtain approximations for the tree length under selection. While Theorem 1 and its corollaries describe the case of genic selection, Theorem 2 and its corollaries deals with other modes of dominance. All proofs are found in Section 5.

Genic selection

Consider a Moran model of size N, where every individual has type either • or ◦, selection is genic, type • is advantageous with selection coefficient α, and mutation is bi-directional. In other words, consider a population of N (haploid) individuals with the following transitions:

1. Every pair of individuals resamples at rate 1; upon such a resampling event, one of the two involved individuals dies, the other one reproduces.

2. Every line is hit by a mutation event from ◦ to • at rate $\theta_\circ > 0$, and by a mutation event from • to ◦ at rate $\theta_\bullet > 0$.

3. Every line of type • places an offspring to a randomly chosen line at rate α.

Mutation leads to an expected change dX of $\theta_\circ(1-X) - X\theta_\bullet = \theta_\circ - X\bar{\theta} = \bar{\theta}(\Theta - X)$ for $\bar{\theta} = \theta_\circ + \theta_\bullet$ and $\Theta = \theta_\circ/\bar{\theta}$ per time dt, and selection of $\alpha X(1-X)dt$. Recall that the frequency X of • in the population follows – in the limit $N \to \infty$ – the SDE

$$dX = \alpha X(1-X)dt + \bar{\theta}(\Theta - X)dt + \sqrt{X(1-X)}dW \tag{1}$$

for some Brownian motion W; see e.g. (5.6) in Ewens (2004).

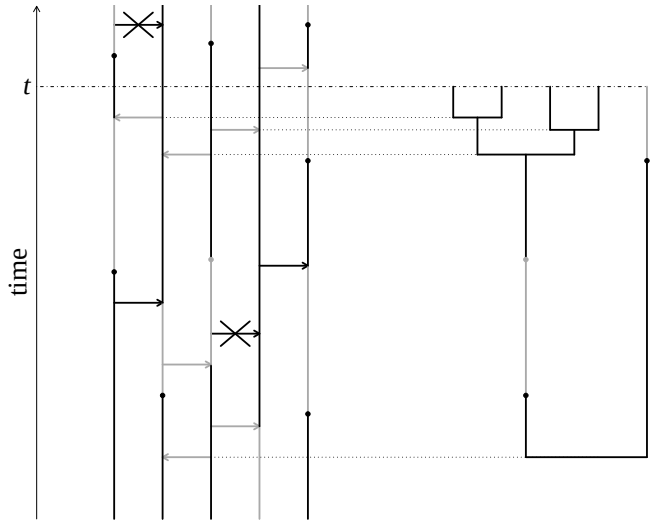

Figure 1: In a population of size N following Moran dynamics with resampling, mutation, and genic selection in equilibrium, it is possible to study genealogical trees as given on the right. Here, the full tree of all individuals in the model is drawn, but it is as well possible to study the tree of a population sample.

In the sequel, we will rely here on the possibility to pick a sample from the Moran population in the large population limit at equilibrium and describe its genealogical tree, which is given by (i) genealogical distances between any pair of individuals, resulting in an ultra-metric tree and (ii) marks on the tree which describe mutation events from • to ◦ or from ◦ to •; see also Figure 1. This possibility is implicitly made by the ancestral selection graph from Neuhauser and Krone (1997), and formally justified by some results obtained in Depperschmidt et al. (2012); precisely, their Theorem 4 states that the genealogical tree under selection has a unique equilibrium.

We will write $\mathbb{P}^\alpha[.]$ for the distribution of genealogical trees under the selection coefficient α and $\mathbb{E}^\alpha[.]$ for the corresponding expectation. In particular, $\mathbb{P}^0[.]$ and $\mathbb{E}^0[.]$ are reserved for neutral evolution, $\alpha = 0$. Within the genealogical tree, we pick a sample of size n and let L_n be the (random) length of its genealogy. We note that in the absence of selection, L_n does not

depend on the mutational mechanism and $L_n \stackrel{d}{=} \sum_{k=2}^{n} kT_k$, where $T_k \sim \exp\binom{k}{2}, k = 2, ..., n$ are the coalescence times in the tree; see e.g. (3.25) in Wakeley (2008). In particular, for $\lambda \geq 0$,

$$f_n := \mathbb{E}^0[e^{-\lambda L_n}] = \prod_{k=2}^{n} \frac{k-1}{k-1+2\lambda} \tag{2}$$

with $f_1 = 1$ since the empty product is defined to be 1. We are now ready to state our first main result, which gives a recursion for an approximation of the Laplace transform of the tree length under selection for small α.

Theorem 1 (Genealogical distances under genic selection)**.** *Let*

$$x_n := \mathbb{E}^\alpha[e^{-\lambda L_n}] - \mathbb{E}^0[e^{-\lambda L_n}].$$

Then, $x_1, x_2, ...$ *satisfy the recursion* $x_1 = 0$ *and*

$$\left(\binom{n}{2} + n\lambda\right) \cdot x_n = \binom{n}{2} \cdot x_{n-1} + \alpha^2 n \cdot a_n + O(\alpha^3), \tag{3}$$

where $a_1, a_2, ...$ *satisfy the recursion* $a_1 = 0$ *and*

$$\left(\binom{n+1}{2} + \bar{\theta} + n\lambda\right) \cdot a_n = \binom{n}{2} \cdot a_{n-1} + \Theta(1-\Theta) \cdot b_n + O(\alpha), \tag{4}$$

where $b_1, b_2, ...$ *satisfy the recursion* $b_1 = 0$ *and*

$$\left(\binom{n+2}{2} + 2\bar{\theta} + n\lambda\right) \cdot b_n = \binom{n}{2} \cdot b_{n-1} + \binom{n}{2} \cdot c_{n-1} + (n-1) \cdot d_n \tag{5}$$

where $c_1, c_2, ...$ *satisfy the recursion* $c_1 = 0$ *and*

$$\left(\binom{n+2}{2} + 2\bar{\theta} + n\lambda\right) \cdot c_n = \binom{n}{2} \cdot c_{n-1} + 2 \cdot e_n + d_n, \tag{6}$$

where $e_1, e_2, ...$ *satisfy a recursion* $e_1 = 0$ *and*

$$\left(\binom{n+1}{2} + 2\bar{\theta} + n\lambda\right) \cdot e_n = \binom{n}{2} \cdot e_{n-1} + d_n \tag{7}$$

and finally – recall (2) *–*

$$d_n = f_{n-1} - f_n - g_{n-1} + g_n \tag{8}$$

with $g_1 = 1/(1+2\bar{\theta})$ *and*

$$g_n = \sum_{b=2}^{n} \frac{n+1}{n-1} \frac{1}{\binom{b+1}{2}} \prod_{k=2}^{b-1} \frac{\binom{k}{2}}{\binom{k}{2} + k\lambda} \prod_{k=b}^{n} \frac{\binom{k}{2}}{\binom{k}{2} + k\lambda + 2\bar{\theta}}. \tag{9}$$

Remark 1 (Interpretations)**.** In the proof, we will see that the quantities $a_n, b_n, c_n, ...$ do have interpretations within the Moran model. If the tree length of the genealogy of a sample of individuals $1, ..., n$ is denoted L_n, the genealogical distance of individuals i and j is R_{ij}, and U_i is the type of individual i (either • of ●), these are

$$\begin{aligned}
a_n &:= \frac{1}{\alpha}\mathbb{E}^{\alpha}[e^{-\lambda L_n}(1_{U_1=\bullet} - 1_{U_{n+1}=\bullet})],\\
b_n &:= \mathbb{E}^0[e^{-\lambda L_n}((n-1)e^{-\bar{\theta}R_{12}} - 2ne^{-\bar{\theta}R_{1,n+1}} + (n+1)e^{-\bar{\theta}R_{n+1,n+2}})],\\
c_n &:= \mathbb{E}^0[e^{-\lambda L_n}(e^{-\bar{\theta}R_{12}} - 2e^{-\bar{\theta}R_{1,n+1}} + e^{-\bar{\theta}R_{n+1,n+2}})],\\
d_n &:= \mathbb{E}^0[(e^{-\lambda L_{n-1}} - e^{-\lambda L_n})(1 - e^{-\bar{\theta}R_{12}})],\\
e_n &:= \mathbb{E}^0[e^{-\lambda L_n}(e^{-\bar{\theta}R_{12}} - e^{-\bar{\theta}R_{1,n+1}})],\\
g_n &:= \mathbb{E}^0[e^{-\lambda L_n}e^{-\bar{\theta}R_{12}}].
\end{aligned} \tag{10}$$

Moreover, in Theorem 2, another quantity will arise, which is

$$h_n = e_n - c_n = \mathbb{E}^0[e^{-\lambda L_n}(e^{-\bar{\theta}R_{1,n+1}} - e^{-\bar{\theta}R_{n+1,n+2}})]. \tag{11}$$

We note that, from these definitions, clearly $a_1 = b_1 = c_1 = e_1 = 0$. The initial value d_2 is given through the initial condition f_1, as well as f_2, g_1 and g_2.

Remark 2 (Comparing neutral and selective genealogies)**.**

1. Note that for $\alpha = 0$, (3) gives precisely (2). Moreover, there is no linear term in α in the recursion (3). This finding is reminiscent of Theorem 4.26 in Krone and Neuhauser (1997) and Theorem 5 in Depperschmidt et al. (2012), but we note that for other models of dominance, a linear term arises; see Theorem 2.

2. Let us compare tree lengths under neutrality and under selection qualitatively. Crucially, the quantity d_n as given in (10) is positive. As consequences, by the recursions, e_n from (7) is positive, c_n from (6) is positive, b_n from (5) is positive, and a_n from (4) is positive. The effect is that x_n for small α is larger than under neutrality, i.e. $\mathbb{E}^{\alpha}[e^{-\lambda L_n}] > \mathbb{E}^0[e^{-\lambda L_n}]$ for small α, which implies that genealogical trees are generally shorter (in the so-called Laplace-transform-order) under selection. In particular, we have shown the intuitive result that expected tree lengths are shorter under selection.

3. While x_n, a_n are quantities within the selected genealogies, all other quantities can be computed under neutrality, $\alpha = 0$. However, if one would like to obtain finer results, i.e. specify the $O(\alpha^3)$-term in (3), more quantities within selected genealogies would have to be computed. In principle, this is straight-forward using our approach of the proof of Theorem 1.

Remark 3 (Solving the recursions)**.** All recursions for $x_n, a_n, b_n, c_n, e_n, h_n$ are of the form

$$\mu_n = \gamma_n \cdot \mu_{n-1} + \nu_n$$

with $\mu_1 = 0$ and can readily be solved by writing

$$\begin{aligned}\mu_n &= \nu_n + \gamma_n \cdot (\nu_{n-1} + \gamma_{n-1} \cdot (\nu_{n-2} + \gamma_{n-2} \cdot (\cdots \nu_2 + \gamma_2 \cdot 0))) \\ &= \sum_{k=2}^{n} \nu_k \prod_{m=k+1}^{n} \gamma_m\end{aligned}$$

with $\prod_{\emptyset} := 1$.

Since we can directly obtain expected tree lengths from the Laplace transforms in Theorem 1, we obtain also a recursion for expected tree lengths.

Corollary 4 (Expected tree length under genic selection)**.** *Let*

$$\widetilde{x}_n := \mathbb{E}^0[L_n] - \mathbb{E}^\alpha[L_n].$$

Then, $\widetilde{x}_1, \widetilde{x}_2, \ldots$ *satisfy the recursion* $\widetilde{x}_1 = 0$ *and*

$$\binom{n}{2} \cdot \widetilde{x}_n = \binom{n}{2} \cdot \widetilde{x}_{n-1} + \alpha^2 n \cdot \widetilde{a}_n + O(\alpha^3), \qquad n = 2, 3, \ldots$$

where $\widetilde{a}_1, \widetilde{a}_2, \ldots$ *satisfy the recursion* $\widetilde{a}_1 = 0$ *and*

$$\left(\binom{n+1}{2} + \bar{\theta}\right) \cdot \widetilde{a}_n = \binom{n}{2} \cdot \widetilde{a}_{n-1} + \Theta(1-\Theta) \cdot \widetilde{b}_n + O(\alpha), \qquad n = 2, 3, \ldots$$

where $\widetilde{b}_1, \widetilde{b}_2, \ldots$ *satisfy the recursion* $\widetilde{b}_1 = 0$ *and*

$$\left(\binom{n+2}{2} + 2\theta\right) \cdot \widetilde{b}_n = \binom{n}{2} \cdot \widetilde{b}_{n-1} + \binom{n}{2} \cdot \widetilde{c}_{n-1} + (n-1) \cdot \widetilde{d}_n$$

where $\widetilde{c}_1, \widetilde{c}_2, \ldots$ *satisfy the recursion* $\widetilde{c}_1 = 0$ *and*

$$\left(\binom{n+2}{2} + 2\bar{\theta}\right) \cdot \widetilde{c}_n = \binom{n}{2} \cdot \widetilde{c}_{n-1} + 2 \cdot \widetilde{e}_n + \widetilde{d}_n,$$

where $\widetilde{e}_1, \widetilde{e}_2, \ldots$ *satisfy a recursion* $\widetilde{e}_1 = 0$ *and*

$$\left(\binom{n+1}{2} + 2\bar{\theta}\right) \cdot \widetilde{e}_n = \binom{n}{2} \cdot \widetilde{e}_{n-1} + \widetilde{d}_n$$

and finally

$$\widetilde{d}_n = \frac{2}{n-1} - \widetilde{g}_{n-1} + \widetilde{g}_n$$

with $\widetilde{g}_1 = 0$ *and*

$$\widetilde{g}_n = \frac{n+1}{n-1}\left(\sum_{b=2}^{n} \frac{1}{\binom{b+1}{2}}\left(\prod_{\ell=b}^{n} \frac{\binom{\ell}{2}}{\binom{\ell}{2} + 2\bar{\theta}}\right)\left(\sum_{k=2}^{b-1} \frac{2}{k-1} + \sum_{k=b}^{n} \frac{k}{\binom{k}{2} + 2\bar{\theta}}\right)\right) \tag{12}$$

The following result, the special case $n = 2$, was already obtained in Theorem 5 of Depperschmidt et al. (2012).

Corollary 5 (Genealogical distance of two individuals under genic selection)**.**

$$\mathbb{E}^{\alpha}[e^{-\lambda L_2}] = \frac{1}{1+2\lambda}\Big(1 + 8\alpha^2\lambda\frac{\Theta(1-\Theta)\bar{\theta}(1+\lambda+\bar{\theta})}{(1+2\bar{\theta})(1+2\lambda)(1+2\lambda+2\bar{\theta})(3+\bar{\theta}+2\lambda)(3+\bar{\theta}+\lambda)}\Big) + O(\alpha^3),$$

$$\mathbb{E}^{\alpha}[L_2] = 2 - 8\alpha^2\frac{\Theta(1-\Theta)\bar{\theta}(1+\bar{\theta})}{(1+2\bar{\theta})^2(3+\bar{\theta})^2} + O(\alpha^3).$$

Other modes of dominance

In a diploid population, (1) only models the frequency of • correctly if selection is genic, i.e. if the selective advantage of an individual which is homozygous for • is twice the advantage of a heterozygote. For other modes of dominance, we have to introduce a dominance coefficient $h \in (-\infty, \infty)$ and replace the selective events in the Moran model by the following transitions:

3'. Let X be the frequency of • in the population. Every line of type • places an offspring to a randomly chosen line at rate $\alpha(X + h(1-X))$. Every line of type ◦ places an offspring to a randomly chosen line at rate $\alpha h X$.

Note that 3'. is best understood by assuming that every line picks a random partner and if the pair is a heterozygote, it has fitness advantage αh, and if it is homozygous for •, it has fitness advantage α. (Here, we have assumed that $h \geq 0$, but some modifications of 3'. also allow for $h < 0$.) The expected effect of 3'. on X is then $\alpha X(1-X)(X + h(1-X)) - \alpha X(1-X)hX = \alpha X(1-X)(h-(1-2h)X)$ and the frequency of • follows – in the limit $N \to \infty$ – the SDE

$$dX = \alpha X(1-X)(h-(1-2h)X)dt + \bar{\theta}(\Theta - X)dt + \sqrt{X(1-X)}dW. \tag{13}$$

We will write $\mathbb{P}^{\alpha,h}[.]$ for the distribution of genealogical trees and allele frequencies under this scenario, and $\mathbb{E}^{\alpha,h}[.]$ for the corresponding expectation. With this notation, we have $\mathbb{P}^{\alpha}[.] = \mathbb{P}^{2\alpha,1/2}[.]$. We note that $h = 0$ means a positively selected recessive allele, while $h = 1$ refers to a dominant selectively favoured allele. Again, we obtain an approximation of the Laplace-transform of the tree length of a sample of size n.

Theorem 2 (Genealogical distances under any form of dominance)**.** *Let $h \in (-\infty, \infty)$ and*

$$y_n := \mathbb{E}^{\alpha,h}[e^{-\lambda L_n}] - \mathbb{E}^{0}[e^{-\lambda L_n}].$$

Then, $y_1, y_2, \ldots$ satisfy the recursion $y_1 = 0$ and

$$\left(\binom{n}{2} + n\lambda\right)\cdot y_n = \binom{n}{2}\cdot y_{n-1} + \alpha n(1-2h)\Theta(1-\Theta)\cdot h_n + O(\alpha^2),$$

where $h_1, h_2, \ldots$ satisfy the recursion $h_1 = 0$ and

$$\left(\binom{n+2}{2} + 2\bar{\theta} + n\lambda\right)\cdot h_n = \binom{n}{2}\cdot h_{n-1} + (n-1)\cdot e_n, \tag{14}$$

and e_n was given in Theorem 1.

Remark 6 (Comparing genealogies)**.** 1. Most interestingly, neutral trees differ from trees under genic selection only in order α^2, whereas the difference is in order α for other forms of dominance. While this may be counter-intuitive at first sight, it can be easily explained. Note that the model actually does not change if we replace α by $-\alpha$ and h by $1-h$ at the same time. By doing so, we just interchange the roles of allele • and ◦. For $h = 1/2$, this means that our results have to be identical for α and $-\alpha$, leading to a vanishing linear term in (3). For $h \neq 1/2$, this symmetry does not have to hold, leading to a linear term in α.

2. Similar to our reasoning in Remark 2.2, the sign of h_n in the recursion for y_n determines if tree lengths are shorter or longer under selection. We see that the behaviour changes at $h = 1/2$. By construction, h_n is positive, so if $h < 1/2$, y_n is positive as well and we see that trees are shorter under selection (in the Laplace-transform order). If $h > 1/2$, the reverse is true and trees are longer under selection. This result is not surprising for over-dominant selection, $h > 1$, since the advantage of the heterozygote leads to maintenance of heterozygosity or balancing selection, which in turn is known to produce longer genealogical trees.

Corollary 7 (Expected tree length under any form of dominance)**.** *Let $h \in (-\infty, \infty)$ and*

$$\widetilde{y}_n := \mathbb{E}^0[L_n] - \mathbb{E}^{\alpha,h}[L_n].$$

Then, $\widetilde{y}_1, \widetilde{y}_2, \ldots$ satisfy the recursion $\widetilde{y}_1 = 0$ and

$$\binom{n}{2} \cdot \widetilde{y}_n = \binom{n}{2} \cdot \widetilde{y}_{n-1} + \alpha n(1-2h)\Theta(1-\Theta) \cdot \widetilde{h}_n + O(\alpha^2),$$

where $\widetilde{h}_1, \widetilde{h}_2, \ldots$ satisfy the recursion $\widetilde{h}_1 = 0$ and

$$\left(\binom{n+2}{2} + 2\bar{\theta}\right) \cdot \widetilde{h}_n = \binom{n}{2} \cdot \widetilde{h}_{n-1} + (n-1) \cdot \widetilde{e}_n, \tag{15}$$

and $\widetilde{e}_n$ was given in Corollary 4.

Corollary 8 (Genealogical distance of two individuals under any form of dominance)**.**

$$\mathbb{E}^\alpha[e^{-\lambda L_2}] = \frac{1}{1+2\lambda}\left(1 + 8\alpha\lambda \frac{(1-2h)\Theta(1-\Theta)\bar{\theta}(1+\lambda+\bar{\theta})}{(1+2\bar{\theta})(1+2\lambda)(1+2\lambda+2\bar{\theta})(3+2\bar{\theta}+2\lambda)(3+\bar{\theta}+\lambda)}\right) + O(\alpha^3),$$

$$\mathbb{E}^\alpha[L_2] = 2 - 8\alpha(1-2h)\frac{\Theta(1-\Theta)\bar{\theta}(1+\bar{\theta})}{(1+2\bar{\theta})^2(3+2\bar{\theta})(3+\bar{\theta})} + O(\alpha^2).$$

3 Discussion

A fundamental question in population genetics is: How does selection affect genealogies of a sample of individuals? We have added to this question an analysis of tree lengths under low levels of selection, both for genic selection and for other modes of dominance. While our results are only given through recursions, these give valuable insights. Collecting previously stated interpretations of our results:

- The selection coefficient enters the change in tree length linearly for $h \neq 1/2$, but only quadratically for genic selection ($h = 1/2$); see Remark 6.1.
- The tree length are shorter under genic selection; see Remark 2.2. For other modes of dominance, tree lengths are shorter only for $h < 1/2$, but longer for $h > 1/2$; see Remark 6.2.

In addition, from Theorems 1 and 2, it is possible to study the effect in large samples. Recall that under neutrality,

$$\tfrac{1}{2}(L_n - 2\log n) \overset{n\to\infty}{\Longrightarrow} Z,$$

where Z is Gumbel distributed (see p. 255 of Wiuf and Hein, 1999). In particular, for large n,

$$\mathbb{E}^0[L_n] \approx 2\log n + 2\gamma_e,$$

where $\gamma_e \approx 0.57$ is the Euler-Mascheroni constant. We show in the appendix that the change of L_n under low levels of selection relative to neutral evolution is $O(1)$, precisely

$$\lim_{\alpha\to 0} \frac{\mathbb{E}^0[L_n] - \mathbb{E}^\alpha[L_n]}{\alpha^2} = O(1), \qquad \lim_{\alpha\to 0} \frac{\mathbb{E}^0[L_n] - \mathbb{E}^{\alpha,h}[L_n]}{\alpha} = O(1) \qquad \text{as } n \to \infty. \tag{16}$$

This then shows that the order of magnitude in the change due to selection is much smaller than the length of the full tree for large samples. Note that this finding is in line with Przeworski et al. (1999), where simulations of the ancestral selection graph are used to find that the overall tree shape under selection is not too different from neutrality for low levels of selection.

In order to get more quantitative insights, we have numerically solved the recursions and plotted the effect on the tree length for various scenarios. Figure 2(A) analyses the effect of genic selection in large samples (i.e. $n = 50$). Interestingly, there is some mutation rate $\bar{\theta} \approx 0.5$, which gives the largest effect. This is clear since very little mutation implies almost no change in the genealogy relative to neutrality since almost always only the beneficial type is present in the population, and very high mutation rate implies that selection is virtually inefficient, leading to nearly neutral trees. Moreover, we can see here that $\Theta(1-\Theta)$ enters the recursion for the change in tree length only linearly. In Figure 2(B), we display the change in tree length for $h = 0$. Since $1 - 2h$ enters the recursion only linearly, the graph looks qualitatively the same for other dominance coefficients.

In principle, the approach we use here is comparable to the ancestral selection graph in the sense that all events happening within the ASG are also implemented in our construction. However, within the ASG, when a splitting event occurs, it is not clear which of the two lines is the true ancestor, so bot lines are followed. As a consequence, in any case the ASG *looks* longer than a neutral coalescent due to the splitting events, and only when the ASG is pruned to become the right genealogical tree, the tree length can be computed. In our approach, the information, which we need within a splitting event is different, since only the type of the additional line, and the type of an individual within the sample is needed.

The approach we use here to study genealogies under selection can be used for other statistics than the total tree length. In principle, every quantity of a sample tree can be described. The reason why we chose the tree length is its simple structure due to coalescence events: If two lines

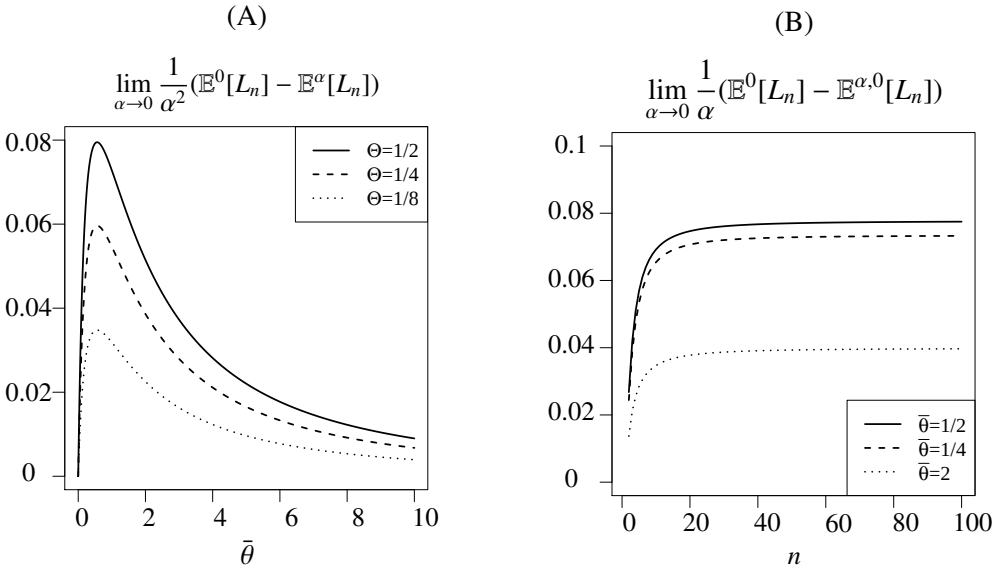

Figure 2: Using the recursions from Corollaries 4 and 7, we see differences in expected tree length. (A) For genic selection and large samples, the effect changes with the total mutation rate $\bar{\theta}$ and is linear in $\Theta(1-\Theta)$. (B) Plot of the change in total tree length for small values of α with $h = 0$, dependent on the sample size, and three parameters of $\bar{\theta}$.

in a sample of size n coalesce, all that remains is a tree of $n-1$ individuals, already implying the recursive structure for the tree lengths which is apparent in Theorems 1 and 2. In addition, in order to describe the effect of selection, we rely on a description of the sample which was already used in the mathematics literature for the so-called Fleming-Viot process (which generalizes the Wright-Fisher diffusion); see e.g. Etheridge (2001) and Depperschmidt et al. (2012). In principle, our approach can be extended e.g. to include more than two alleles, population structure, recombination etc. However, probably one would still have to find a recursive structure, which will often be feasible only in the case of weak selection.

4 Preliminaries for the proofs

We here present the construction of a tree-valued process in a nutshell, leaving out various technical details. All details of the construction are given in Depperschmidt et al. (2012).

Any genealogical tree is uniquely given by all genealogical distances of pairs on (haploid) individuals. So, in order to describe the evolution of genealogical trees, it suffices to describe the evolution of all pairwise distances. We also note that the tree length which we consider in all our results, is a function of pairwise distances. (See e.g. Section 8 of Depperschmidt et al. (2012).) Consider a sample of size n taken from the Moran model at time t as described in Section 2, and let $R(t) := (R_{ij}(t))_{i<j}$ be the pairwise genealogical distances (note that $R_{ij} = R_{ji}$), and $U(t) := (U_1(t), ..., U_n(t))$ the allelic types, either • or ●. We will consider some smooth, bounded

function $\Phi : \mathbb{R}_+^{\binom{n}{2}} \times \{\bullet, \circ\}^n \to \mathbb{R}$ and are going to describe the change in $\mathbb{E}^\alpha[\Phi(R(t), U(t))]$ due to the evolution of the Moran model. We have to take into account several mechanisms:

0. Growth of the tree: During times when no event happens, all genealogical distances grow deterministically and linearly (with speed 2). In time dt, the change is

$$\mathbb{E}^\alpha[\Phi(R(t) + 2dt, U(t)) - \Phi(R(t), U(t))] = \mathbb{E}^\alpha\Big[2\sum_{i<j}\frac{\partial}{\partial r_{ij}}\Phi(R(t), U(t))\Big]dt. \tag{17}$$

1. Resampling: If a pair of the N individuals within the Moran model resamples, there is either none, one, or two of them within the sample tree of size n. If none, the sample tree is not affected. If one, and this one reproduces, the sample tree is not affected as well. If one, and this one is replaced by the individual outside of the sample, the effect is the same as if we would have picked the other individual to begin with. Since in the $\mathbb{E}^\alpha[...]$, we average over all possibilities which samples of size n we take, there is also no resulting effect. If two, i reproduces and j dies, say, the effect is that distances to individual j are replaced by distances to individual i, and the new type of individual j is the type of i. Since all pairs within the sample resample at rate 1, the change in time dt is

$$\mathbb{E}^\alpha\Big[\sum_{i<j}\Phi(\theta_{ij}(R(t)), \theta_{ij}(U(t))) - \Phi(R(t), U(t))\Big]dt \tag{18}$$

with

$$(\theta_{ij}(R(t)))_{k\ell} = \begin{cases} R_{k\ell}(t), & \text{if } k, \ell \neq j, \\ R_{i\ell}(t), & \text{if } k = j, \\ R_{ki}(t), & \text{if } \ell = j, \end{cases} \qquad (\theta(U(t)))_k = \begin{cases} U_k(t), & \text{if } k \neq j, \\ U_i(t), & \text{if } k = j. \end{cases}$$

2. Mutation: We note that the mutational mechanism can also be described by saying that every line in the Moran model mutates at rate $\bar{\theta}$, with outcome $\bullet$ with probability Θ and outcome $\circ$ with probability $1-\Theta$. Since mutation only affects the alleles of the individuals in the sample, we have in time dt

$$\bar{\theta} \cdot \mathbb{E}^\alpha\Big[\sum_i \Theta\Phi(R(t), U^{i,\bullet}(t)) + (1-\Theta)\Phi(R(t), U^{i,\circ}(t)) - \Phi(R(t), U(t))\Big]dt \tag{19}$$

with

$$(U^{i,\bullet}(t))_k = \begin{cases} U_k(t), & \text{if } k \neq i, \\ \bullet, & \text{if } k = i. \end{cases}$$

and analogously for $U^{i,\circ}$.

3. Selection: By the dynamics of the Moran model, we have to accept that selective events depend on the type of individuals. We say that the kth individual has fitness $\alpha\chi_k :=$

$\alpha\chi(U_k)$, and χ is the fitness function. For genic selection and other modes of dominance, the fitness function is given by

$$\chi_k = \begin{cases} 1, & \text{if } U_k = \bullet, \\ 0, & \text{if } U_k = \cdot, \end{cases} \qquad \chi_k = \begin{cases} 1, & U_k = U_m = \bullet, \\ h, & U_k \neq U_m, \\ 0, & U_k = U_m = \cdot, \end{cases}$$

respectively. Here, for other modes of dominance, m is some randomly picked (haploid) individual. As for resampling, selective events occur in a pair of one individual i giving birth and the other, individual j, dying, as given through the function θ_{ij} from above. Consequently, we find that in time dt, since $n \ll N$, and all individuals outside the sample bring the same effect, for genic selection

$$\begin{aligned} &\frac{\alpha}{N}\mathbb{E}^\alpha\Big[\sum_{i<j}\chi_i(\Phi(\theta_{ij}(R(t)),\theta_{ij}(U(t))) - \Phi(R(t),U(t)))\Big]dt \\ &\quad\approx \alpha\mathbb{E}^\alpha\Big[\sum_{j=1}^{n}\chi_{n+1}(\Phi(\theta_{n+1,j}(R(t)),\theta_{n+1,j}(U(t))) - \Phi(R(t),U(t)))\Big]dt \\ &\quad= \alpha\mathbb{E}^\alpha\Big[\sum_{j=1}^{n}\chi_j\Phi(R(t),U(t)) - \chi_{n+1}\Phi(R(t),U(t)))\Big]dt. \end{aligned} \tag{20}$$

In the $\approx$ (which is exact in the limit $N \to \infty$), we have used that the effect of selective events within the sample can be neglected. For the last equality, we have used that we can permute sampling orders in $\mathbb{E}^\alpha[.]$ since all genealogies are sampled with the same probability and by changing individuals j and $n+1$ in the sample in the first term.

We will now focus on the function, for some $0 \le i \le n$, $j \ge 0$,

$$\Phi^n_{ij}(t) := \Phi^n_{ij}(R(t),U(t)) := e^{-\lambda L_n(t)}1(U_1(t) = \cdots = U_i(t) = U_{n+1}(t) = \cdots = U_{n+j}(t) = \bullet), \tag{21}$$

and note that the goal of Theorems 1 and 2 is to approximate $\mathbb{E}^\alpha[\Phi^n_{00}(\infty)] = \mathbb{E}^\alpha[e^{-\lambda L_n}]$. The following lemma is an application of the general theory from 0.-3. above.

Lemma 9. *For $n \ge 2$, and with the convention that $\Phi^n_{-1,j} = \Phi^n_{i,-1} = 0$,*

$$\begin{aligned} \frac{d}{dt}\mathbb{E}^\alpha[\Phi^n_{ij}(t)] &= -n\lambda\cdot\mathbb{E}^\alpha[\Phi^n_{ij}(t)] + \binom{i}{2}\cdot\mathbb{E}^\alpha[\Phi^{n-1}_{i-1,j}(t)] + \left(\binom{n}{2} - \binom{i}{2}\right)\cdot\mathbb{E}^\alpha[\Phi^{n-1}_{i,j}(t)] \\ &+ (n-i)j\cdot\mathbb{E}^\alpha[\Phi^n_{i+1,j-1}(t)] + ij\cdot\mathbb{E}^\alpha[\Phi^n_{i,j-1}(t)] + \binom{j}{2}\cdot\mathbb{E}^\alpha[\Phi^n_{i,j-1}(t)] \\ &+ i\theta_{\cdot}\cdot\mathbb{E}^\alpha[\Phi^n_{i-1,j}(t)] + j\theta_{\cdot}\cdot\mathbb{E}^\alpha[\Phi^n_{i,j-1}(t)] \\ &+ \alpha\cdot\mathbb{E}^\alpha[(n-i)\Phi^n_{i+1,j} - (n+j)\Phi^n_{i,j+1}] - \left(\binom{n+j}{2} + (i+j)\bar\theta - \alpha(i+j)\right)\cdot\mathbb{E}^\alpha[\Phi^n_{i,j}(t)] \end{aligned} \tag{22}$$

for genic selection, and the last two terms change to

$$\begin{aligned}\alpha\cdot\mathbb{E}^\alpha[&((i+j)\Phi^n_{i,j+1}-(n+2i+3j)h\Phi^1_{i,j+1}+(n-i)h\Phi^n_{i+1,j}\\&+(n-i)(1-2h)\Phi^n_{i+1,j+1}-(n+j)(1-2h)\Phi^n_{i,j+2}]\\&-\left(\binom{n+j}{2}+(i+j)\bar\theta-\alpha h(i+j)\right)\cdot\mathbb{E}^\alpha[\Phi^n_{i,j}(t)]\end{aligned}\tag{23}$$

for other modes of dominance.

Remark 10. Simple algebra shows that the α-terms in (22) and (23) agree for $h = 1/2$. For future reference, note that for $i = j = 0$, the α-term in (23) gives

$$\alpha\cdot\mathbb{E}^\alpha[nh(\Phi^n_{10}-\Phi^{n+1}_{01})+n(1-2h)(\Phi^{n+1}_{11}-\Phi^{n+2}_{02})].$$

Proof of Lemma 9. The effect of tree growth on $\mathbb{E}^\alpha[\Phi^n_{ij}]$ is that the tree growth by ndt in time dt, i.e. in time dt

$$-n\lambda\cdot\mathbb{E}^\alpha[\Phi^n_{ij}(t)].$$

Let $I = \{1,...,i\}, H = \{i+1,...,n\}$ and $J = \{n+1,...,n+j\}$. For resampling, we distinguish between events among I, events among $I\cup H$, with at most one partner within I, events with one partner within I and the second among J, events with one partner in H and the second among J, and events with two partners in J. Only if two among $I\cup H$ coalesce, n decreases. This gives

$$\begin{aligned}&\binom{i}{2}(\mathbb{E}^\alpha[\Phi^{n-1}_{i-1,j}(t)]-\mathbb{E}^\alpha[\Phi^n_{i,j}(t)])\\&\quad+\left(\binom{n}{2}-\binom{i}{2}\right)(\mathbb{E}^\alpha[\Phi^{n-1}_{i,j}(t)]-\mathbb{E}^\alpha[\Phi^n_{i,j}(t)])\\&\quad+ij(\mathbb{E}^\alpha[\Phi^n_{i+1,j-1}(t)]-\mathbb{E}^\alpha[\Phi^n_{i,j}(t)])\\&\quad+(n-i)j(\mathbb{E}^\alpha[\Phi^n_{i,j-1}(t)]-\mathbb{E}^\alpha[\Phi^n_{i,j}(t)])\\&\quad+\binom{j}{2}(\mathbb{E}^\alpha[\Phi^n_{i,j-1}(t)]-\mathbb{E}^\alpha[\Phi^n_{i,j}(t)])\end{aligned}$$

For mutation, we note that for $h\in H$, $\Phi^n_{ij}(R,U^{h,\bullet}) = \Phi^n_{ij}(R,U^{h,\circ}) = \Phi^n_{ij}$, therefore the effect of mutation is

$$i\cdot\mathbb{E}^\alpha[\theta_\circ\cdot(\Phi^n_{i-1,j}-\Phi^n_{ij})]+j\cdot\mathbb{E}^\alpha[\theta_\circ\cdot(\Phi^n_{i,j-1}-\Phi^n_{ij})]-(i+j)\cdot\mathbb{E}^\alpha[\theta_\bullet\cdot\Phi^n_{ij}].$$

Last, for selection, we have to distinguish the cases of genic selection and other modes of dominance. For genic selection, we have that $\chi_k = 1(U_k=\bullet)$ and we note that for $i\in I$ and $j\in J$, $\Phi^n_{ij}\chi_i = \Phi^n_{ij}\chi_j = \Phi^n_{ij}$, therefore the effect is

$$\alpha\cdot\mathbb{E}^\alpha[(i+j)(\Phi^n_{ij}-\Phi^n_{i,j+1})+(n-i)(\Phi^n_{i+1,j}-\Phi^n_{i,j+1})]$$

For other modes of dominance,

$$\begin{aligned}\chi_k&=1(U_k=U_{n+j+1}=\bullet)+h(1(U_k=\bullet)+1(U_{n+j+1}=\bullet)-2\cdot1(U_k=U_{n+j+1}=\bullet))\\&=(1-2h)\cdot1(U_k=U_{n+j+1}=\bullet)+h(1(U_k=\bullet)+1(U_{n+j+1}=\bullet)).\end{aligned}$$

Therefore, the effect of selection is here

$$\begin{aligned}
&\alpha\cdot\mathbb{E}^{\alpha}[i((1-2h)\Phi^n_{i,j+1}+h(\Phi^n_{ij}+\Phi^n_{i,j+1}))]\\
&\quad+\mathbb{E}^{\alpha}[(n-i)\cdot((1-2h)\Phi^n_{i+1,j+1}+h(\Phi^n_{i+1,j}+\Phi^n_{i,j+1})]\\
&\quad+\mathbb{E}^{\alpha}[j\cdot((1-2h)\Phi^n_{i,j+1}+h(\Phi^n_{ij}+\Phi^n_{i,j+1})]\\
&\quad-\mathbb{E}^{\alpha}[(n+j)\cdot((1-2h)\Phi^n_{i,j+2}+2h\Phi^n_{i,j+1}]\\
&=\alpha\cdot\mathbb{E}^{\alpha}[((i+j)h\Phi^n_{ij}+i(1-h)+(n-i)h+j(1-h)-2(n+j)h)\Phi^n_{i,j+1}\\
&\qquad+(n-i)h\Phi^n_{i+1,j}+(n-i)(1-2h)\Phi^n_{i+1,j+1}-(n+j)(1-2h)\Phi^n_{i,j+2}].
\end{aligned}$$

□

The next result is collected from Theorem 4 and Lemma 8.1 in Depperschmidt et al. (2012).

Lemma 11. *The process (R,U) of genealogical distances and types, has a unique equilibrium under $\mathbb{P}^{\alpha,h}$. This equilibrium is described by $\frac{d}{dt}\mathbb{E}^{\alpha,h}[\Phi(R(t),U(t))]=0$ for all possible Φ. Moreover, for this equilibrium, denoted by $(R(\infty),U(\infty))$, satisfies*

$$\mathbb{E}^{\alpha}[\Phi^n_{ij}]=\mathbb{E}^{0}[\Phi^n_{ij}]+O(\alpha)\ \text{as}\ \alpha\to 0.$$

5 Proof of Theorems 1 and 2

Proof of Theorem 1. To begin, we note that the quantities as defined in Remark 1 for $n=1$ – since $L_1=0$ – are given by $a_1=b_1=c_1=e_1=0$. Moreover, we note that $\mathbb{E}^0[e^{-\lambda L_n}(1(U_1=\bullet)-1(U_{n+1}=\bullet))]=0$ since the mutational history of single lines, leading to U_1 and U_{n+1} are independent of the genealogy for $\alpha=0$ and therefore, from Lemma 11, we see that $\widetilde{a}_n:=\alpha a_n=O(\alpha)$. For the recursion on x_n, we write – from Lemma 9 –

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}^{\alpha}[e^{-\lambda L_n(t)}]=\frac{d}{dt}\mathbb{E}^{\alpha}[\Phi^n_{00}(t)]&=-n\lambda\mathbb{E}^{\alpha}[e^{-\lambda L_n(t)}]+\binom{n}{2}\cdot(\mathbb{E}^{\alpha}[e^{-\lambda L_{n-1}(t)}]-\mathbb{E}^{\alpha}[e^{-\lambda L_n(t)}])\\
&\quad+\alpha n\cdot(\mathbb{E}^{\alpha}[e^{-\lambda L_n(t)}(1(U_1=\bullet)-1(U_{n+1}=\bullet))].
\end{aligned}$$

In equilibrium, the right hand side must equal 0, and solving for $\mathbb{E}^{\alpha}[e^{-\lambda L_n}]$ gives (3), where a_n is given in (10). For the recursion on a_n, we write with Lemma 9

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}^{\alpha}[e^{-\lambda L_n(t)}(1(U_1=\bullet)-1(U_{n+1}=\bullet))]&=\frac{d}{dt}\mathbb{E}^{\alpha}[\Phi^n_{10}(t)-\Phi^n_{01}(t)]\\
=\binom{n}{2}\cdot\mathbb{E}^{\alpha}[\Phi^{n-1}_{10}(t)-\Phi^{n-1}_{01}(t)]&+\alpha\cdot\mathbb{E}^{\alpha}[(n-1)\Phi^n_{20}(t)-2n\Phi^n_{11}(t))+(n+1)\Phi^n_{02}(t)]\\
&-\left(\binom{n+1}{2}+\bar{\theta}+n\lambda-\alpha\right)\cdot\mathbb{E}^{\alpha}[\Phi^n_{10}(t)-\Phi^n_{01}(t)].
\end{aligned}$$

In equilibrium, both sides must be 0, and dividing both sides by α gives a recursion for a_n, but we still have to show that the last term is $b_n+O(\alpha)$. First, we can replace $\mathbb{E}^{\alpha}[.]$ by $\mathbb{E}^{0}[.]$ in

this expression, since we are making an error of order α at most. Then, for $\alpha = 0$, i.e. neutral evolution, we note that two individuals $k \neq \ell$, which have genealogical distance $R_{k\ell}$, both have type $\bullet$ in either of two cases: (i) there is no mutation on the path between k, ℓ in the genealogy, and their joint ancestor has type $\bullet$; (ii) there is a mutation on the path between k, ℓ, and both mutational events determining the types of k, ℓ give the type $\bullet$. For $\alpha = 0$, the mutational process is independent of coalescence events, hence we find the probabilities $(1 - e^{-\bar{\theta}R_{k\ell}})\Theta$ and $(1 - e^{-\bar{\theta}R_{k\ell}})\Theta^2$ in cases (i) and (ii), respectively. Hence, for any $k, \ell = 1, ..., n+2$ and $k \neq \ell$

$$\begin{aligned}\mathbb{E}^0[e^{-\lambda L_n}, U_k = U_\ell = \bullet] &= \mathbb{E}^0\big[e^{-\lambda L_n}(e^{-\bar{\theta}R_{k\ell}}\Theta + (1 - e^{-\bar{\theta}R_{k\ell}})\Theta^2)\big] \\ &= \mathbb{E}^0\big[e^{-\lambda L_n}\big] + \Theta(1-\Theta)\mathbb{E}\big[e^{-\lambda L_n}e^{-\bar{\theta}R_{k\ell}}\big].\end{aligned}$$

Hence, we obtain that

$$\begin{aligned}&\mathbb{E}^0[(n-1)\Phi^n_{20}(t) - 2n\Phi^n_{11}(t)) + (n+1)\Phi^n_{02}(t)] \\ &\qquad = \Theta(1-\Theta)\cdot\mathbb{E}^0[e^{-\lambda L_n}((n-1)e^{-\bar{\theta}R_{12}} - 2ne^{-\bar{\theta}R_{1,n+1}} + (n+1)e^{-\bar{\theta}R_{n+1,n+2}})],\end{aligned}$$

which proves (4), where b_n is given as in (10). For the remaining recursions, we always work with $\alpha = 0$. In order to obtain a recursion for b_n, consider a coalescent with $n+2$ lines and distinguish the following cases for the first step:

1. Coalescence of lines among the first n lines, except for lines 1,2 (rate $\binom{n}{2} - 1$);
2. Coalescence of lines 1,2 (rate 1);
3. Coalescence of lines $n+1$ and 1 (rate 1);
4. Coalescence of lines $n+1$ and one of $2, ..., n$ (rate $n-1$);
5. Coalescence of lines $n+1$ and $n+2$ (rate 1);
6. Coalescence of lines $n+2$ and one of $1, ..., n$ (rate n);

Recalling $T_{n+2} \sim \exp\binom{n+2}{2}$, we write by a first-step decomposition

$$
\begin{aligned}
&\frac{1}{\mathbb{E}[e^{-(n\lambda+2\bar{\theta})T_{n+2}}]}\binom{n+2}{2}\cdot\mathbb{E}[e^{-\lambda L_n}((n-1)e^{-\bar{\theta}R_{12}}-2ne^{-\bar{\theta}R_{1,n+1}}+(n+1)e^{-\bar{\theta}R_{n+1,n+2}})]\\
&=\left(\binom{n}{2}-1\right)\cdot\mathbb{E}[e^{-\lambda L_{n-1}}((n-1)e^{-\bar{\theta}R_{12}}-2ne^{-\bar{\theta}R_{1,n}}+(n+1)e^{-\bar{\theta}R_{n,n+1}})]\\
&\quad+1\cdot\mathbb{E}[e^{-\lambda L_{n-1}}((n-1)-2ne^{-\bar{\theta}R_{1,n}}+(n+1)e^{-\bar{\theta}R_{n,n+1}})]\\
&\quad+1\cdot\mathbb{E}[e^{-\lambda L_n}((n-1)e^{-\bar{\theta}R_{12}}-2n+(n+1)e^{-\bar{\theta}R_{1,n+1}})]\\
&\quad+(n-1)\cdot\mathbb{E}[e^{-\lambda L_n}((n-1)e^{-\bar{\theta}R_{12}}-2ne^{-\bar{\theta}R_{12}}+(n+1)e^{-\bar{\theta}R_{1,n+1}})]\\
&\quad+1\cdot\mathbb{E}[e^{-\lambda L_n}((n-1)e^{-\bar{\theta}R_{12}}-2ne^{-\bar{\theta}R_{1,n+1}}+(n+1))]\\
&\quad+n\cdot\mathbb{E}[e^{-\lambda L_n}((n-1)e^{-\bar{\theta}R_{12}}-2ne^{-\bar{\theta}R_{1,n+1}}+(n+1)e^{-\bar{\theta}R_{1,n+1}})]\\
&=\binom{n}{2}\cdot\mathbb{E}[e^{-\lambda L_{n-1}}((n-1)e^{-\bar{\theta}R_{12}}-2ne^{-\bar{\theta}R_{1,n}}+(n+1)e^{-\bar{\theta}R_{n,n+1}})]\\
&\quad+(n-1)\cdot\mathbb{E}[e^{-\lambda L_{n-1}}(1-e^{-\bar{\theta}R_{12}})]\\
&\quad+\underbrace{((n-1-(n-1)(n+1)+n-1+n(n-1))}_{=n-1}\cdot\mathbb{E}[e^{-\lambda L_n}e^{-\bar{\theta}R_{12}}]\\
&\quad+\underbrace{(n+1+(n-1)(n+1)-2n-n(n-1))}_{=0}\cdot\mathbb{E}[e^{-\lambda L_n}e^{-\bar{\theta}R_{1,n+1}}]\\
&\quad-(n-1)\cdot\mathbb{E}[e^{-\lambda L_n}]\\
&=(n-1)\cdot((\mathbb{E}[(e^{-\lambda L_{n-1}}-e^{-\lambda L_n})(1-e^{-\bar{\theta}R_{12}})])\\
&\quad+\binom{n}{2}\cdot\mathbb{E}[e^{-\lambda L_{n-1}}((n-1)e^{-\bar{\theta}R_{12}}-2ne^{-\bar{\theta}R_{1,n}}+(n+1)e^{-\bar{\theta}R_{n,n+1}})].
\end{aligned}
$$

This shows (5). For c_n, we use the same coalescent, and by distinguishing the six cases, we write

$$\begin{aligned}
&\frac{1}{\mathbb{E}[e^{-(n\lambda+2\bar\theta)T_{n+2}}]}\binom{n+2}{2}\mathbb{E}[e^{-\lambda L_n}(e^{-\bar\theta R_{12}} - 2e^{-\bar\theta R_{1,n+1}} + e^{-\bar\theta R_{n+1,n+2}})]\\
&= \left(\binom{n}{2} - 1\right)\cdot\mathbb{E}[e^{-\lambda L_{n-1}}(e^{-\bar\theta R_{12}} - 2e^{-\bar\theta R_{1,n}} + e^{-\bar\theta R_{n,n+1}})] + 1\cdot\mathbb{E}[e^{-\lambda L_{n-1}}(1 - 2e^{-\bar\theta R_{1,n}} + e^{-\bar\theta R_{n,n+1}})]\\
&\quad + 1\cdot\mathbb{E}[e^{-\lambda L_n}(e^{-\bar\theta R_{12}} - 2 + e^{-\bar\theta R_{1,n+1}})] + (n-1)\cdot\mathbb{E}[e^{-\lambda L_n}(e^{-\bar\theta R_{12}} - 2e^{-\bar\theta R_{12}} + e^{-\bar\theta R_{1,n+1}})]\\
&\quad + 1\cdot\mathbb{E}[e^{-\lambda L_n}(e^{-\bar\theta R_{12}} - 2e^{-\bar\theta R_{1,n+1}} + 1)] + n\cdot\mathbb{E}[e^{-\lambda L_n}(e^{-\bar\theta R_{12}} - 2e^{-\bar\theta R_{1,n+1}} + e^{-\bar\theta R_{1,n+1}})]\\
&= \binom{n}{2}\cdot\mathbb{E}[e^{-\lambda L_{n-1}}(e^{-\bar\theta R_{12}} - 2e^{-\bar\theta R_{1,n}} + e^{-\bar\theta R_{n,n+1}})] + \mathbb{E}[e^{-\lambda L_{n-1}}(1 - e^{-\bar\theta R_{12}})]\\
&\quad + \underbrace{(1-n+1+1+n)}_{=3}\cdot\mathbb{E}[e^{-\lambda L_n}e^{-\bar\theta R_{12}}] + \underbrace{(1+n-1-2-n)}_{=-2}\cdot\mathbb{E}[e^{-\lambda L_n}e^{-\bar\theta R_{1,n+1}}] - \mathbb{E}[e^{-\lambda L_{n-1}}]\\
&= \mathbb{E}[(e^{-\lambda L_{n-1}} - e^{-\lambda L_n})(1 - e^{-\bar\theta R_{12}})] + 2\cdot\mathbb{E}[e^{-\lambda L_n}(e^{-\bar\theta R_{12}} - e^{-\bar\theta R_{1,n+1}})]\\
&\qquad + \binom{n}{2}\cdot\mathbb{E}[e^{-\lambda L_{n-1}}(e^{-\bar\theta R_{12}} - 2e^{-\bar\theta R_{1,n}} + e^{-\bar\theta R_{n,n+1}})],
\end{aligned}$$

which shows (6). For d_n, let $B \in \{2, ..., n\}$ be the number of lines in a coalescent starting with n lines, just before lines 1 and 2 coalesce. Then,

$$\begin{aligned}
\mathbb{P}(B = b) &= \frac{1}{\binom{b}{2}}\prod_{k=b+1}^{n}\frac{\binom{k}{2}-1}{\binom{k}{2}} = \frac{2}{b(b-1)}\prod_{k=b+1}^{n}\frac{(k+1)(k-2)}{k(k-1)} = \frac{2}{b(b-1)}\frac{(n+1)(b-1)}{(b+1)(n-1)}\\
&= \frac{n+1}{n-1}\frac{1}{\binom{b+1}{2}}.
\end{aligned}$$

Then,

$$\begin{aligned}
g_n &:= \mathbb{E}^0[e^{-\lambda L_n}e^{-\bar\theta R_{12}}] = \mathbb{E}^0[e^{-\lambda\sum_{k=2}^n kT_k}e^{-\bar\theta\sum_{k=B}^n 2T_k}]\\
&= \sum_{b=2}^{n}\frac{n+1}{n-1}\frac{1}{\binom{b+1}{2}}\prod_{k=2}^{b-1}\frac{\binom{k}{2}}{\binom{k}{2}+k\lambda}\prod_{k=b}^{n}\frac{\binom{k}{2}}{\binom{k}{2}+k\lambda+2\bar\theta}
\end{aligned}$$

and we see that

$$d_n = f_{n-1} - f_n - g_{n-1} + g_n.$$

Finally, for e_n, we again use a recursion. Consider a coalescent with $n+1$ lines and make a first-step-analysis. In this first step, we distinguish four cases:

1. Coalescence of lines 1 or 2 with one of $3, ..., n$; rate $\binom{n}{2} - 1$
2. Coalescence of lines 1 and 2; rate 1
3. Coalescence of lines $n+1$ and 1; rate 1

4. Coalescence of lines $n+1$ and one of $2, ..., n$; rate $n-1$

Hence,

$$\begin{aligned}
&\frac{1}{\mathbb{E}[e^{-(n\lambda+2\bar\theta)T_{n+1}}]}\binom{n+1}{2}\cdot\mathbb{E}[e^{-\lambda L_n}(e^{-\bar\theta R_{12}}-e^{-\bar\theta R_{1,n+1}})]\\
&\qquad=\left(\binom{n}{2}-1\right)\cdot\mathbb{E}[e^{-\lambda L_{n-1}}(e^{-\bar\theta R_{12}}-e^{-\bar\theta R_{1,n}})]+1\cdot\mathbb{E}[e^{-\lambda L_{n-1}}(1-e^{-\bar\theta R_{1,n}})]\\
&\qquad\quad+1\cdot\mathbb{E}[e^{-\lambda L_n}(e^{-\bar\theta R_{12}}-1)]+(n-1)\cdot\mathbb{E}[e^{-\lambda L_n}(e^{-\bar\theta R_{12}}-e^{-\bar\theta R_{12}})]\\
&\qquad=\binom{n}{2}\cdot\mathbb{E}[e^{-\lambda L_{n-1}}(e^{-\bar\theta R_{12}}-e^{-\bar\theta R_{1,n}})]+\mathbb{E}[(e^{-\lambda L_{n-1}}-e^{-\lambda L_n})(1-e^{-\bar\theta R_{12}})].
\end{aligned}$$

This shows (7). □

Proof of Corollary 4. We have to compute $\widetilde{x}_n := \frac{\partial}{\partial\lambda}x_n|_{\lambda=0}$. A close inspection of the recursions for x_n reveals that (i) x_n is a sum of products, and in each summand, some factor d_k enters and (ii) $d_k = O(\lambda)$ for small λ for all k, which is best seen from (10). As a consequence, we can compute the derivative with respect to λ at $\lambda = 0$ in each summand which enters x_n by taking the derivative of d_k with respect to to λ and set $\lambda = 0$ in all other factors. Summing up, we have the same recursions as in Theorem 1 with (i) $\lambda = 0$ in all terms except d_n, and (ii) replace d_n with the derivative according to λ at $\lambda = 0$. This gives the recursions as given in the corollary and

$$\widetilde{d}_n = \frac{\partial}{\partial\lambda}f_n - f_{n-1} - g_{n-1} + g_n\Big|_{\lambda=0} = \frac{2}{n-1} + \widetilde{g}_{n-1} - \widetilde{g}_n$$

with

$$\begin{aligned}
\widetilde{g}_n &= -\frac{\partial}{\partial\lambda}\sum_{b=2}^{n}\frac{n+1}{n-1}\frac{1}{\binom{b+1}{2}}\prod_{k=2}^{b-1}\frac{\binom{k}{2}}{\binom{k}{2}+k\lambda}\prod_{\ell=b}^{n}\frac{\binom{\ell}{2}}{\binom{\ell}{2}+\ell\lambda+2\bar\theta}\Big|_{\lambda=0}\\
&= \frac{n+1}{n-1}\Big(\sum_{b=2}^{n}\sum_{k=2}^{b-1}\frac{1}{\binom{b+1}{2}}\frac{2}{k-1}\prod_{\ell=b}^{n}\frac{\binom{\ell}{2}}{\binom{\ell}{2}+2\bar\theta}+\sum_{b=2}^{n}\sum_{k=b}^{n}\frac{1}{\binom{b+1}{2}}\frac{\binom{k}{2}k}{\left(\binom{k}{2}+2\bar\theta\right)^2}\prod_{\substack{\ell=b\\ \ell\neq k}}^{n}\frac{\binom{\ell}{2}}{\binom{\ell}{2}+2\bar\theta}\Big)
\end{aligned}$$

and the result follows. □

Proof of Corollary 5. Applying Theorem 1, we get

$$\begin{aligned}
g_2 &= \frac{1}{1+2\lambda+2\bar\theta},\\
d_2 &= 1-\frac{1}{1+2\lambda}-\frac{1}{1+2\bar\theta}+\frac{1}{1+2\lambda+2\bar\theta} = \frac{2\lambda}{1+2\lambda}-\frac{2\lambda}{(1+2\bar\theta)(1+2\lambda+2\bar\theta)}\\
&= \frac{2\lambda(1+2\lambda+4\bar\theta+4\lambda\bar\theta+4\bar\theta^2-1+2\lambda)}{(1+2\bar\theta)(1+2\lambda)(1+2\lambda+2\bar\theta)} = \frac{8\lambda\bar\theta(1+\lambda+\bar\theta)}{(1+2\bar\theta)(1+2\lambda)(1+2\lambda+2\bar\theta)},\\
a_2 &= \frac{\Theta(1-\Theta)}{3+\bar\theta+2\lambda}\cdot b_2 = \frac{\Theta(1-\Theta)}{3+\bar\theta+2\lambda}\cdot\frac{1}{6+2\bar\theta+2\lambda}d_2.
\end{aligned}$$

This gives the first assertion. The second follows since a_2 is $O(\lambda)$ and the derivative with respect to λ at $\lambda = 0$ is easily computed. □

Proof of Theorem 2. For the selection operator, applied to the Laplace transform of the tree length, we have

$$\begin{aligned}
&\alpha n\mathbb{E}[e^{-\lambda L_n}(\chi(u_1,u_{n+1})-\chi(u_{n+1},u_{n+2}))]\\
&\quad=\alpha n\mathbb{E}[\mathbb{E}[e^{-\lambda L_n}\chi(u_1,u_{n+1})|R_{1,n+1}]-\mathbb{E}[e^{-\lambda L_n}\chi(u_{n+1},u_{n+2})|R_{n+1,n+2}]]\\
&\quad=\alpha n\mathbb{E}[e^{-\lambda L_n}e^{-\bar\theta R_{1,n+1}}\Theta+(1-e^{-\bar\theta R_{1,n+1}})\Theta^2+2h(1-e^{-\bar\theta R_{1,n+1}})\Theta(1-\Theta)\\
&\qquad\quad-e^{-\lambda L_n}e^{-\bar\theta R_{n+1,n+2}}\Theta+(1-e^{-\bar\theta R_{n+1,n+2}})\Theta^2+2h(1-e^{-\bar\theta R_{n+1,n+2}})\Theta(1-\Theta)]\\
&\quad=\alpha n\Theta(1-\Theta)(1-2h)\mathbb{E}[e^{-\lambda L_n}(e^{-\bar\theta R_{1,n+1}}-e^{-\bar\theta R_{n+1,n+2}})]
\end{aligned}$$

Now, we want to get a recursion for the last term. Consider a coalescent with $n+2$ lines and distinguish the following cases:

1. Coalescence of lines among the first n lines (rate $\binom{n}{2}$);
2. Coalescence of lines $n+1$ and 1 (rate 1);
3. Coalescence of lines $n+1$ and one of $2,...,n$ (rate $n-1$);
4. Coalescence of lines $n+1$ and $n+2$ (rate 1);
5. Coalescence of lines $n+2$ and one of $1,...,n$ (rate n).

We then get

$$\begin{aligned}
&\frac{1}{\mathbb{E}[e^{-(n\lambda+2\bar\theta)T_{n+2}}]}\binom{n+2}{2}\mathbb{E}[e^{-\lambda L_n}(e^{-\bar\theta R_{1,n+1}}-e^{-\bar\theta R_{n+1,n+2}})]\\
&=\binom{n}{2}\cdot\mathbb{E}[e^{-\lambda L_{n-1}}(e^{-\bar\theta R_{1,n}}-e^{-\bar\theta R_{n,n+1}})]\\
&\quad+1\cdot\mathbb{E}[e^{-\lambda L_n}(1-e^{-\bar\theta R_{1,n+1}})]\\
&\quad+(n-1)\cdot\mathbb{E}[e^{-\lambda L_n}(e^{-\bar\theta R_{12}}-e^{-\bar\theta R_{1,n+1}})]\\
&\quad+1\cdot\mathbb{E}[e^{-\lambda L_n}(e^{-\bar\theta R_{1,n+1}}-1)]\\
&\quad+n\cdot\mathbb{E}[e^{-\lambda L_n}(e^{-\bar\theta R_{1,n+1}}-e^{-\bar\theta R_{1,n+1}})]\\
&=\binom{n}{2}\cdot\mathbb{E}[e^{-\lambda L_{n-1}}(e^{-\bar\theta R_{1,n}}-e^{-\bar\theta R_{n,n+1}})]+(n-1)\cdot\mathbb{E}[e^{-\lambda L_n}(e^{-\bar\theta R_{12}}-e^{-\bar\theta R_{1,n+1}})].
\end{aligned}$$

□

Proof of Corollary 7. The proof is basically the same as for Corollary 4: The recursions for y_n is a sum of products, where each factor comes with a factor d_n, and $d_n = O(\lambda)$. Therefore, the derivative according to λ at $\lambda = 0$ is performed by taking derivatives only of d_n and setting $\lambda = 0$ in all other instances. □

Proof of Corollary 8. Applying Theorem 2, we get

$$e_2=\frac{d_2}{3+2\bar\theta+2\lambda},\qquad h_2=\frac{e_2}{6+2\bar\theta+2\lambda},$$

and with d_2 from the proof of Corollary 5, the result follows. □

A Proof of (16)

For the first assertion, we have that $\widetilde{d}_n \sim \frac{1}{n}$, where $a_n \sim b_n$ if $0 < \liminf_{n\to\infty} \frac{a_n}{b_n} \leq \limsup_{n\to\infty} \frac{a_n}{b_n} < \infty$. From Remark 3, we can solve the recursions for $\widetilde{e}_n, \widetilde{c}_n, \widetilde{b}_n$ and $\widetilde{a}_n$, which all are of the form

$$\left(\binom{n}{2} + \kappa n + \rho\right) \cdot \mu_n = \binom{n}{2} \cdot \mu_{n-1} + \nu_n.$$

We want to find the behaviour of μ_n for large n. Hence,

$$\mu_n \sim \sum_{k=2}^{n} \frac{\nu_k}{k^2} \exp\Big(- \sum_{m=k+1}^{n} \log\Big(1 + \frac{2\kappa}{m}\Big)\Big) \sim \sum_{k=2}^{n} \frac{\nu_k}{k^2} \exp\Big(- \sum_{m=k+1}^{n} \frac{2\kappa}{m}\Big) \sim \sum_{k=2}^{n} \frac{\nu_k}{k^2}\Big(\frac{k}{n}\Big)^{2\kappa}.$$

Since $\widetilde{d}_n \sim 1/n$ and the recursion for $\widetilde{e}_n$ comes with $\kappa = 1$, we find that

$$\widetilde{e}_n \sim \sum_{k=2}^{n} \frac{1}{k^3}\Big(\frac{k}{n}\Big)^2 \sim \frac{\log n}{n^2}.$$

Next, the recursion for $\widetilde{c}_n$ comes with $\kappa = 2$, and $2 \cdot \widetilde{e}_n + \widetilde{d}_n \sim \widetilde{d}_n$, so $\widetilde{c}_n \sim \frac{1}{n^2}$. In the recursion for $\widetilde{b}_n$, we have $\kappa = 2$, so $\widetilde{b}_n \sim \frac{1}{n}$. In the recursion for $\widetilde{a}_n$, we have $\kappa = 1$, so $\widetilde{a}_n \sim \frac{1}{n}\log n$. Finally, for $\widetilde{x}_n$, we write

$$\lim_{\alpha\to 0} \frac{\widetilde{x}_n}{\alpha^2} = \lim_{\alpha\to 0} \frac{1}{\alpha^2} \sum_{k=2}^{n} \widetilde{x}_k - \widetilde{x}_{k-1} \sim \sum_{k=2}^{n} \frac{\log k}{k^2} \sim 1,$$

the first assertion of (16).

Similarly, in order to study the effect under other modes of dominance, we have that the recursion for $\widetilde{h}_n$ comes with $\kappa = 2$, therefore $\widetilde{h}_n \sim \frac{\log n}{n^2}$. Then,

$$\lim_{\alpha\to 0} \frac{\widetilde{y}_n}{\alpha} = \lim_{\alpha\to 0} \frac{1}{\alpha} \sum_{k=2}^{n} \widetilde{y}_k - \widetilde{y}_{k-1} \sim \sum_{k=2}^{n} \widetilde{h}_k \sim 1,$$

which gives the second assertion of (16).

Acknowledgements

This research was supported by the DFG priority program SPP 1590, and in particular through grant Pf-672/6-1 to PP.

References

Barton, N., A. Etheridge, and A. Sturm (2004). Coalescence in a random background. *Annals of Applied Probability 14, no. 2*, 754–785.

Coop, G. and R. C. Griffiths (2004). Ancestral inference on gene trees under selection. *Theo. Pop. Biol. 66*(3), 219–232.

Depperschmidt, A., A. Greven, and P. Pfaffelhuber (2012). Tree-valued Fleming–Viot dynamics with mutation and selection. *Annals of Applied Probability 22*(6), 2560–2615.

Etheridge, A. (2001). *An introduction to superprocesses*. American Mathematical Society.

Ewens, W. (2004). *Mathematical PopulationGenetics. I. Theoretical introduction. Second edition*. Springer.

Ewing, G. and J. Hermisson (2010). Msms: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics 26*(16), 2064–2065.

Hudson, R. (1983). Properties of a neutral allele model with intragenic recombination. *Theo. Pop. Biol. 23*, 183–201.

Kaplan, N., T. Darden, and R. Hudson (1988). The coalescent process in models with selection. *Genetics 120*, 819–829.

Kern, A. D. and D. R. Schrider (2016). Discoal: flexible coalescent simulations with selection. *Bioinformatics 32*(24), 3839–3841.

Kingman, J. (1982). The coalescent. *Stochastic Process. Appl. 13*(3), 235–248.

Krone, S. and C. Neuhauser (1997). Ancestral processes with selection. *Theo. Pop. Biol. 51*, 210–237.

Neuhauser, C. and S. Krone (1997). The genealogy of samples in models with selection. *Genetics 154*, 519–534.

Przeworski, M., B. Charlesworth, and J. D. Wall (1999). Genealogies and weak purifying selection. *Molecular biology and evolution 16*(2), 246–252.

Taylor, J. (2007). The common ancestor process for a Wright-Fisher diffusion. *Electron. J. Probab. 12*, 808–847.

Wakeley, J. (2008). *Coalescent Theory: An Introduction*. Roberts & Company.

Wakeley, J. (2010). Natural selection and coalescent theory. In *Evolution since Darwin: The First 150 Years*, pp. 119–149. Sunderland, MA: Sinauer and Associates.

Wiuf, C. and J. Hein (1999). Recombination as a point process along sequences. *Theo. Pop. Biol. 55*, 248–259.