October 8, 2015

The multivariate Wright-Fisher process with mutation: Moment-based analysis and inference using a hierarchical Beta model

Asger Hobolth[1] and Jukka Siren[2]

Running title: Moment-based analysis of the multivariate Wright-Fisher model.

Abstract

We consider the diffusion approximation of the multivariate Wright-Fisher process with mutation. Analytically tractable formulae for the first- and second-order moments of the allele frequency distribution are derived, and the moments are subsequently used to better understand key population genetics parameters and modelling frameworks. In particular we investigate the behaviour of the expected homozygosity (the probability that two randomly sampled genes are identical) in the transient and stationary phases, and how appropriate the Dirichlet distribution is for modelling the allele frequency distribution at different evolutionary time scales. We find that the Dirichlet distribution is adequate for the pure drift model (no mutations allowed), but the distribution is not sufficiently flexible for more general mutation models. We suggest a new hierarchical Beta distribution for the allele frequencies in the Wright-Fisher process with a mutation model on the nucleotide level that distinguishes between transitions and transversions.

Key words: Allele frequency, diffusion, Dirichlet model, hierarchical Beta, moments, multivariate Wright-Fisher.

1 Introduction

Present day data sets for studying genetic variation within and between species often consists of millions of markers and hundreds to thousands of individuals. The huge number of individuals makes tree-based analyses (e.g. based on phylogenetics or coalescent theory) difficult because the number of possible trees increases very fast with the number of individuals. This difficulty is pronounced when studying closely related species, where incomplete lineage sorting or deep coalescence events can distort phylogenetic analyses (Maddison, 1997). The discrepancy between species trees and gene trees can be taken

[1]Corresponding author. Bioinformatics Research Center, Aarhus University, asger@birc.au.dk. **Mailing address:** Bioinformatics Research Center, Aarhus University, C.F. Møllers Alle, Building 1110, DK-8000 Aarhus C, Denmark. **Phone:** (+45)-8715-5573. **Fax:** (+45)-8715-4102.
[2]Department of Biosciences, University of Helsinki, jukka.p.siren@helsinki.fi

into account by using multispecies coalescence methods (Degnan and Rosenberg, 2009; Heled and Drummond, 2010). However, this more detailed framework is computationally more challenging because the unknown gene trees need to be marginalized out from the model in order to carry out species tree inference. In some special cases the gene trees can be marginalized out using dynamic programming techniques (e.g. Bryant, Bouckaert, Felsenstein *et al.*, 2012), but in general it is neccessary to perform large-scale Monte Carlo simulations to do the integration (Heled and Drummond, 2010).

An attractive alternative to tree-based methodology is to model the allele frequencies over time in terms of a diffusion process, which is derived as an infinite population limit of the Wright-Fisher model. Unfortunately the transition density for the diffusion process corresponding to the basic Wright-Fisher model with mutation remains unknown; the solution to the Fokker-Planck equation is not available (Ewens, 2004, Chapter 5). Numerical approximations have been proposed to approximate the transition density, but they are limited to a small number of populations or species due to computational complexity (e.g. Gutenkunst, Hernandez, Williamson *et al.* 2009). The numerical solutions also assume that each site has experienced at most one mutation and consequently has at most two alleles, which restricts their usage to closely related samples. For more distantly related samples where multiallelic loci are expected to occur, it is important to generalize to the multivariate case (Jenkins, Mueller and Song, 2014).

An alternative strategy to numerically solve the Fokker-Planck equation is to approximate the transition density by a parametric distribution. This methodology has a long tradition in population genetics and computational phylogenetics starting from the seminal work by Edwards, Cavalli-Sforza and Felsenstein in the 60's and 70's (Edwards and Cavalli-Sforza, 1964; Cavalli-Sforza and Edwards, 1967; Felsenstein, 1973), and continuing to present day (Nicholson *et al.* 2002; Gaggiotti and Foll, 2010; Siren, Marttinen and Corander, 2011; Pickrell and Pritchard, 2012). However, most of the methods have been developed for situations where the time span is sufficiently short to ignore mutations and consider only pure drift. Furthermore, the parametric distributions have been either the Gaussian or Dirichlet distributions.

We derive the first- and second-order moments of the multi-allelic Wright-Fisher process with mutation and use the moments to characterise genetic variation and to fit parametric models. Our approach generalizes the work by Siren (2012) and Siren, Hanage and Corander (2013) to arbitrary mutation models. In the first part of the paper (Section 2) we provide new analytically tractable formulas for the first- and second-order moments of the multivariate Wright-Fisher model with mutation. These new formulas allows us to characterise the expected mean and (co)variance of the frequency of an allele, and in particular we investigate in detail the expected homozygosity (Section 3). Furthermore we demonstrate how our formulas can be used to re-derive previous results for the various general symmetric models considered in Griffiths (1980). We emphasize that our mutation stucture is completely unrestricted.

In the second part of the paper (Section 4) we use the expressions for the means and (co)variances of the allele frequencies to obtain insight into approximate models

for the allele frequency distribution over time. In particular we find that while the Dirichlet model is a suitable approximate model for the allele frequency distribution in the Wright-Fisher process with no mutation (pure drift), it is not appropriate for the Wright-Fisher model with a mutation structure that corresponds to the Kimura model. Instead, we propose a novel hierarchical Beta model for the Wright-Fisher process with Kimura mutations. The paper ends with a brief summary of our main findings, and a discussion of similar methodology.

2 First- and second-order moments in the Wright-Fisher with mutation process

We consider a constant-sized haploid population with N individuals. We denote by $z(m) = (z_1(m), \ldots, z_K(m))$ the row-vector of the number of alleles $1, \ldots, K$ in generation m, and we let U be the $K \times K$ mutation probability matrix such that U_{ij} is the probability for a mutation from allele i to allele j in a generation. The Wright-Fisher model with mutation is then given by the multinomial distribution

$$z(m+1)|z(m) \sim \text{Mult}(N, x(m)U), \tag{1}$$

where $x(m) = z(m)/N$ is the allele frequency in generation m.

We are now in a position to formulate our main result:

Theorem 1. General formulas for the mean and variance in the Wright-Fisher with mutation process
Consider the K-allele Wright-Fisher model with mutation probability matrix U and with initial allele frequency $x(0)$. Define the rate matrix $Q = N(U - I)$. In the diffusion approximation the mean of the allele frequency is given by

$$\text{E}[x(t)|x(0)] = x(0)e^{Qt}, \tag{2}$$

and the variance is given by

$$\text{Var}[x(t)|x(0)] = \int_0^t e^{-s}(e^{Qs})'\text{diag}\{x(0)e^{Q(t-s)}\}(e^{Qs})ds - (e^{Qt})'x(0)'x(0)e^{Qt}(1 - e^{-t}). \tag{3}$$

Here we make use of a slight abuse of notation such that $x(t)$ is the allele frequency distribution in generation Nt.

Despite the huge interest in the Wright-Fisher process we believe the clean formula for the variance is a new result.

Proof. Repeated use of the law of total expectation gives the mean value

$$\text{E}[x(m)] = \text{E}\big[\text{E}[x(m)|x(m-1)]\big] = \text{E}[x(m-1)U] = \text{E}[x(m-1)]U = \cdots = x(0)U^m,$$

where for ease of notation we have omitted the conditioning on $x(0)$. We approximate U^m as follows

$$U^m = U^{tN} = \left[\left\{I + (U - I)\right\}^N\right]^t = \left[\left\{I + Q/N\right\}^N\right]^t \approx (e^Q)^t = e^{Qt},$$

where we scale time as $m = tN$ and define $Q = N(U - I)$. Note that with this definition Q becomes a rate matrix where off-diagonal entries are non-negative and rows sum to zero. Thus we have, with a small abuse of notation,

$$\mathrm{E}[x(t)|x(0)] = x(0)e^{Qt}.$$

The proof of the variance is more involved, but the main idea is to make repeated use of the law of total variance. The proof can be found in Appendix A. □

Very many procedures are available for calculating matrix exponentials (e.g. Moler and Van Loan, 2003), so a numerical calculation of the mean is straight forward. Calculating the variance is more difficult. In Appendix B we provide an analytical expression for the mean and variance in the case of a reversible mutation matrix. The expression is based on an eigenvalue decomposition of the rate matrix.

There is a long tradition for careful investigation of mutation models in phylogenetics (e.g. Felsenstein, 2004, Chapter 13). In this paper we consider in particular the pure drift model ($U = I$; see Corollary 3 and Corollary 5), the Jukes-Cantor model ($U_{ij} = u,\ i \neq j$; see Corollary 4 and Corollary 8), and the symmetric model ($U = U'$; see Theorem 7 and Theorem 9). We give special attention to the Kimura model (Felsenstein, 2004, page 196-200) with $K = 4$ and mutation probability matrix

$$U_{ij} = \begin{cases} \kappa u & \text{if mutation } i \to j \text{ is a transition} \\ u & \text{if mutation } i \to j \text{ is a transversion} \\ 1 - (\kappa + 2)u & \text{if } i = j \end{cases}$$

or, equivalently,

$$Q_{ij} = \begin{cases} N\kappa u & \text{if mutation } i \to j \text{ is a transition} \\ Nu & \text{if mutation } i \to j \text{ is a transversion} \\ -N(\kappa + 2)u & \text{if } i = j. \end{cases} \tag{4}$$

We parameterize the rate matrix using either $\alpha = N\kappa u$ (the rate for a transition) and $\beta = Nu$ (the rate for a transversion), or using $\kappa = \alpha/\beta$ (the ratio of the transition rate and transversion rate) and $\theta = \alpha + 2\beta = N(\kappa + 2)u$ (the mutation rate).

The matrix exponential for the Kimura model is given by

$$e^{Qt} = \frac{1}{4}E + \frac{1}{4}e^{-4\beta t}A + \frac{1}{2}e^{-2(\alpha+\beta)t}B, \tag{5}$$

where E is the matrix with one in every entry,

$$A = \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{pmatrix} \text{ and } B = \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}.$$

In Appendix C we derive, based on Theorem 1, analytical expressions for the Kimura model for any initial frequency. In Figure 1 we show as solid lines the mean and variance for the Kimura model with parameters $\kappa = 10$, $\theta = 1$ and a initial frequency $x(0) = (70, 25, 4, 1)/100$ (in the order A,G,C,T). We note that convergence toward the stationary values is not always monotone and that it may take a long time to converge, which implies that drift is important even with this high mutation rate.

In Figure 1 we also show a fit to the mean and (co)variance structure at various time points. The dashed lines correspond to a fit based on the Dirichlet distribution. The Dirichlet distribution with $K = 4$ categories has 4 parameters. Three of these parameters determine the mean, leaving one parameter to model the (co)variance structure. It is clear from the figure that the Dirichlet distribution is not flexible enough to capture the behaviour of the second-order moments of the allele frequencies. In particular one single parameter is clearly not enough to describe the covariance of the stationary distribution; the covariance at stationarity has two limiting points corresponding to a transition or a transversion.

In Figure 1 we also show a fit based on the hierarchical Beta distribution that we define in Section 4 below. The hierarchical Beta distribution has 6 parameters. Three of the parameters fully determine the mean, and the remaining 3 parameters nicely capture the (co)variance structure of the model (dotted points).

In order to demonstrate the applicability of Theorem 1 we consider the problem of determining the mean and variance in the case where the initial frequency is the uniform distribution. From the limiting values when the evolutionary time gets large we can then obtain the stationary values of the mean, variance, and covariance. In Figure 1 we have included these stationary points as solid horizontal lines.

Corollary 2. Mean and variance in the Kimura model with uniform initial frequency
In the Kimura model with $x(0) = (1, 1, 1, 1)/4 = \mathbf{e}/4$ *the mean is given by*

$$\mathrm{E}[x(t)|x(0) = \mathbf{e}/4] = \mathbf{e}/4,$$

and the variance is given by

$$\begin{aligned} \mathrm{Var}[x(t)|x(0) = \mathbf{e}/4] &= \frac{1}{16}\Big\{\frac{1}{8\beta+1}(1 - e^{-(8\beta+1)t})A + \frac{2}{4(\alpha+\beta)+1}(1 - e^{-(4(\alpha+\beta)+1)t})B\Big\} \\ &\to \frac{1}{16}\Big\{\frac{A}{8\beta+1} + \frac{2B}{4(\alpha+\beta)+1}\Big\}. \end{aligned}$$

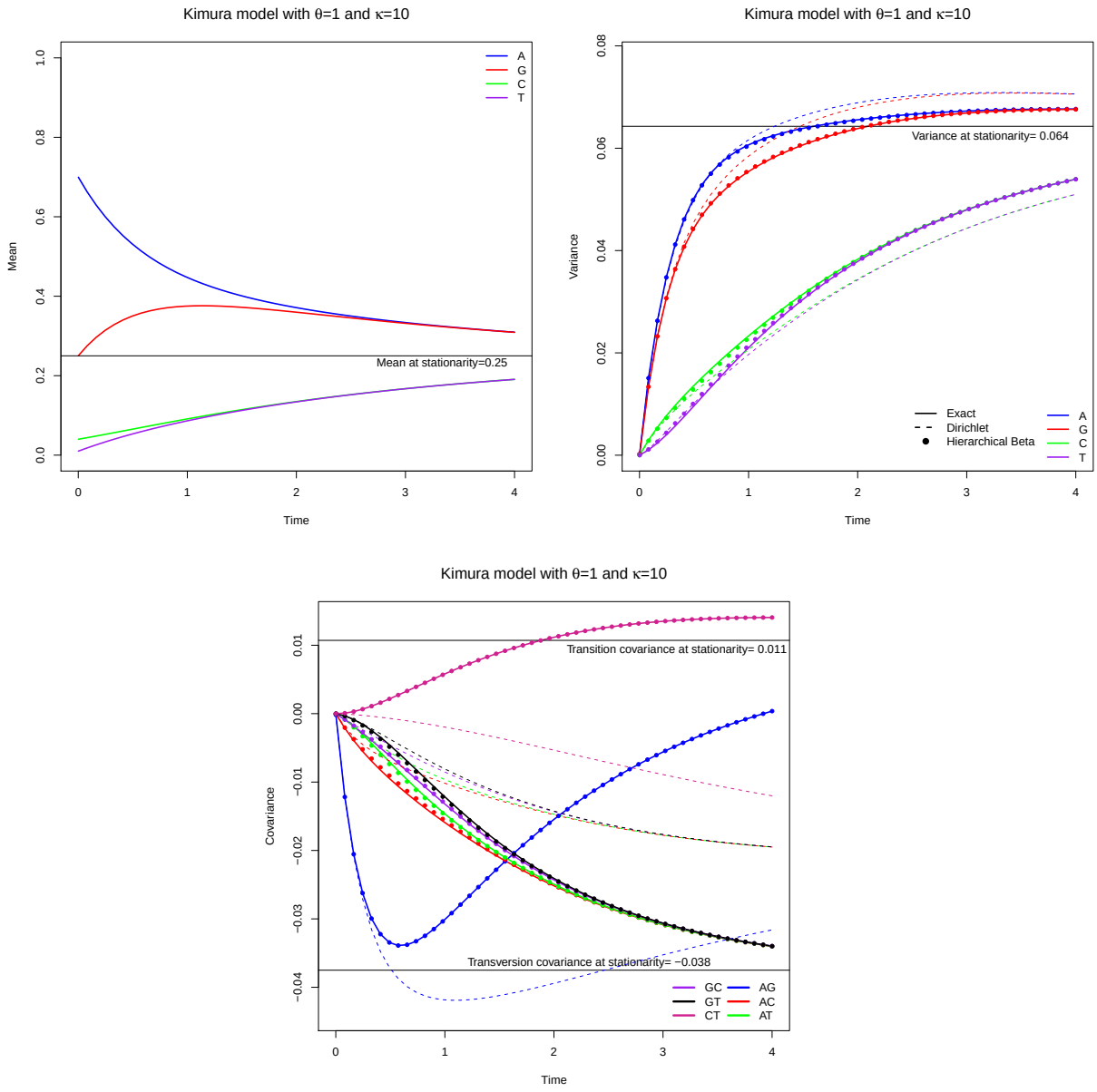

Figure 1: The Kimura model as an illustration of Theorem 1. The initial allele frequencies are $(70, 25, 4, 1)/100$. The three plots show the four means, four variances and six covariances as a function of time. Time is scaled such that one time unit corresponds to N generations where N is the population size. The solid lines are the true functions of the moments. The dashed line is a fit based on the Dirichlet distribution, and the bullets is a fit based on the hierarchical Beta distribution proposed in Section 4. The hierarchical Beta model fits the true behaviour of the means and (co)variances much better than the Dirichlet model. The values at stationarity come from Corollary 2.

Proof. The corollary follows by inserting the matrix exponential (5) in Theorem 1. □

The special case of pure drift also follows easily from Theorem 1:

Corollary 3. Mean and variance in the pure drift model
If $U = I$ we get the mean and variance

$$\begin{aligned} \mathrm{E}[x(t)|x(0)] &= x(0) \\ \mathrm{Var}[x(t)|x(0)] &= (1 - e^{-t})\Big[\mathrm{diag}\{x(0)\} - x(0)'x(0)\Big]. \end{aligned}$$

Proof. In the pure drift case we have $U = I$ and therefore $Q = 0$ and $e^{Qt} = I$. Now the result is an immediate consequence of Theorem 1. □

A natural approximate model for the allele frequency distribution in the pure drift model is to assume the Dirichlet distribution

$$x(t)|x(0) \sim \mathrm{Dir}(\alpha),$$

where $\alpha = (\alpha_1, \ldots, \alpha_K)$ is the vector of K free parameters. The reason is that the mean and variance in the Dirichlet model are

$$\mathrm{E}[x(t)|x(0] = \alpha/\alpha_0 \tag{6}$$

$$\mathrm{Var}[x(t)|x(0)] = \frac{1}{\alpha_0 + 1}\Big[\mathrm{diag}\{\alpha/\alpha_0\} - (\alpha/\alpha_0)'(\alpha/\alpha_0)\Big] \tag{7}$$

where $\alpha_0 = \sum_{k=1}^{K} \alpha_k$, and therefore we have a 1-1 correspondence between $(x(0), t)$ and α, namely $1 - e^{-t} = 1/(\alpha_0 + 1)$ and $x(0) = \alpha/\alpha_0$. Indeed this model is known as the Balding-Nichols model; we refer to Balding and Steele (2015, Section 5.3.2) for more information. We note that the fixation index $F_{\mathrm{ST}} = 1/(\alpha_0 + 1)$, and in the case of pure drift we have $F_{\mathrm{ST}} = 1 - e^{-t}$ (see Balding and Steele, 2015, Section 5.2, regarding the F_{ST} index).

Another special case where we can obtain rather nice explicit analytical results is the Jukes-Cantor model:

Corollary 4. Mean and variance in the Jukes-Cantor model
In the Jukes-Cantor model with rate matrix $Q = (q_{ij})$ given by

$$q_{ij} = \begin{cases} -q & \text{if } i = j \\ q/(K-1) & \text{if } i \neq j, \end{cases}$$

we get the mean

$$\mathrm{E}[x(t)|x(0)] = e^{-\epsilon t/2}(x(0) - \mathbf{e}/K) + \mathbf{e}/K, \tag{8}$$

and variance

$$
\begin{aligned}
&\mathrm{Var}[x(t)|x(0)] = \\
&\Big[I - E/K\Big]/K \frac{1}{1+\epsilon}\Big(1 - e^{-(1+\epsilon)t}\Big) - \\
&(x(0) - \mathbf{e}/K)'(x(0) - \mathbf{e}/K)e^{-\epsilon t}(1 - e^{-t}) + \\
&\Big[\mathrm{diag}(x(0) - \mathbf{e}/K) - (x(0) - \mathbf{e}/K)'\mathbf{e}/K - (\mathbf{e}/K)'(x(0) - \mathbf{e}/K)\Big] \times \\
&e^{-\epsilon t/2} \frac{1}{1+\epsilon/2}\Big(1 - e^{-(1+\epsilon/2)t}\Big).
\end{aligned} \tag{9}
$$

Here $\mathbf{e}$ *is the vector of length* K *with 1 in every entry,* $E = \mathbf{e}'\mathbf{e}$ *is the* $K \times K$ *matrix with 1 in every entry, and* $\epsilon = 2qK/(K-1)$.

Proof. In matrix notation the matrix exponential is given by

$$
e^{Qt} = e^{-\epsilon t/2}(I - E/K) + E/K, \tag{10}
$$

and the expression for the mean follows immediately from Theorem 1. The expression for the variance requires more calculations; the details are provided in Appendix D. □

A few comments are in order. We first note that for large t we have the mean and variance

$$
\mathrm{E}[x(t)|x(0)] = \mathbf{e}/K \;\text{ and }\; \mathrm{Var}[x(t)|x(0)] = \frac{1}{1+\epsilon}\Big[I - E/K\Big]/K,
$$

as expected because the stationary distribution is the Dirichlet distribution with parameter $\epsilon(\mathbf{e}/K)$ (see e.g. Ewens, 2004, page 195).

Second we note from (7) that in order for the Dirichlet model to be appropriate, the variance should be proportional (at least approximately) to the following function of the mean

$$
\begin{aligned}
&\mathrm{diag}\{\mathrm{E}[x(t)]\} - \mathrm{E}[x(t)]'\mathrm{E}[x(t)] = \\
&\mathrm{diag}\{e^{-\epsilon t/2}(x(0) - \mathbf{e}/K) + \mathbf{e}/K\} - \\
&(e^{-\epsilon t/2}(x(0) - \mathbf{e}/K) + \mathbf{e}/K)'(e^{-\epsilon t/2}(x(0) - \mathbf{e}/K) + \mathbf{e}/K) = \\
&(I - E/K)/K - \\
&(x(0) - \mathbf{e}/K)'(x(0) - \mathbf{e}/K)e^{-\epsilon t} + \\
&\Big[\mathrm{diag}(x(0) - \mathbf{e}/K) - (x(0) - \mathbf{e}/K)'\mathbf{e}/K - e'/K(x(0) - \mathbf{e}/K)\Big]e^{-\epsilon t/2}
\end{aligned} \tag{11}
$$

where we used expression (8) for the mean of the fully symmetric model. We observe, as expected from the considerations regarding the pure drift model, that expression (11) is approximately proportional to expression (9) with proportionality constant $(1 - \exp(-t))$ for small ϵ. Furthermore we observe, using that for small t we have the Taylor expansion

$(1-\exp(-(1+a)t))/(1+a) \approx t$, that for small evolutionary distances the two expressions are also approximately proportional with constant of proportionality t and regardless of the mutation rate ϵ.

We conclude that the Dirichlet model is a good approximation of allele frequency distribution in the Wright-Fisher with a Jukes-Cantor mutation model when the mutation rate is small or the evolutionary distance is small. In general, however, the Dirichlet model is not appropriate. For example in Figure 1 we demonstrated that for the Wright-Fisher with Kimura mutations the Dirichlet model has too few parameters to capture the covariance structure. In Section 4 below we propose a new model for the Wright-Fisher with Kimura mutations: The hierarchical Beta model. We demonstrate that the hierarchical Beta model is sufficiently flexible to model allele frequency behaviour for the Kimura mutation model.

Before discussing modelling strategies we demonstrate how the new formulas for the first- and second-order moments can be used to investigate the expected behaviour of homozygosity.

3 Homozygosity in the Wright-Fisher model

Griffiths (1980) provide formulas for the expected behaviour of the homozygosity in the transient and stationary phases. The homozygosity

$$F(t) = \sum_{k=1}^{K} x_k(t)^2 = x(t)x(t)'$$

is the probability of sampling two genes of the same allelic type. The expected homozygosity is a function of the mean and variance

$$\mathrm{E}[F(t)] = \mathrm{trace}(\mathrm{Var}[x(t)]) + \mathrm{E}[x(t)]\mathrm{E}[x(t)]'. \tag{12}$$

Nice analytical expressions for the expected homozygosity are generally not available in the Wright-Fisher with general mutation model, but Griffiths (1980) provided several special cases where analytical expressions are available. In this section we re-derive three main results in Griffiths (1980) using the formulas from Theorem 1. The results all follow easily from Lemma 6 below.

In order to illustrate the results in this section we show in Figure 2 in blue the expected homozygosity as a function of time for the Kimura model with $\kappa = 10$, initial frequencies $(4, 1, 0, 0)/5$ (in light blue) and $(4, 0, 1, 0)/5$ (in dark blue), and three different values of the mutation rate θ. These curves are based on our main Theorem 1 and the formula for homozygosity (12). The Jukes-Cantor model (in purple) is based on Corollary 8 with $\epsilon = 2\theta K/(K-1)$, the curves for the pure drift model (in red) are based on Corollary 5, and the homozygosity at stationarity is calculated from Theorem 7 below. Finally we discuss the curves in green in connection with Theorem 9 below.

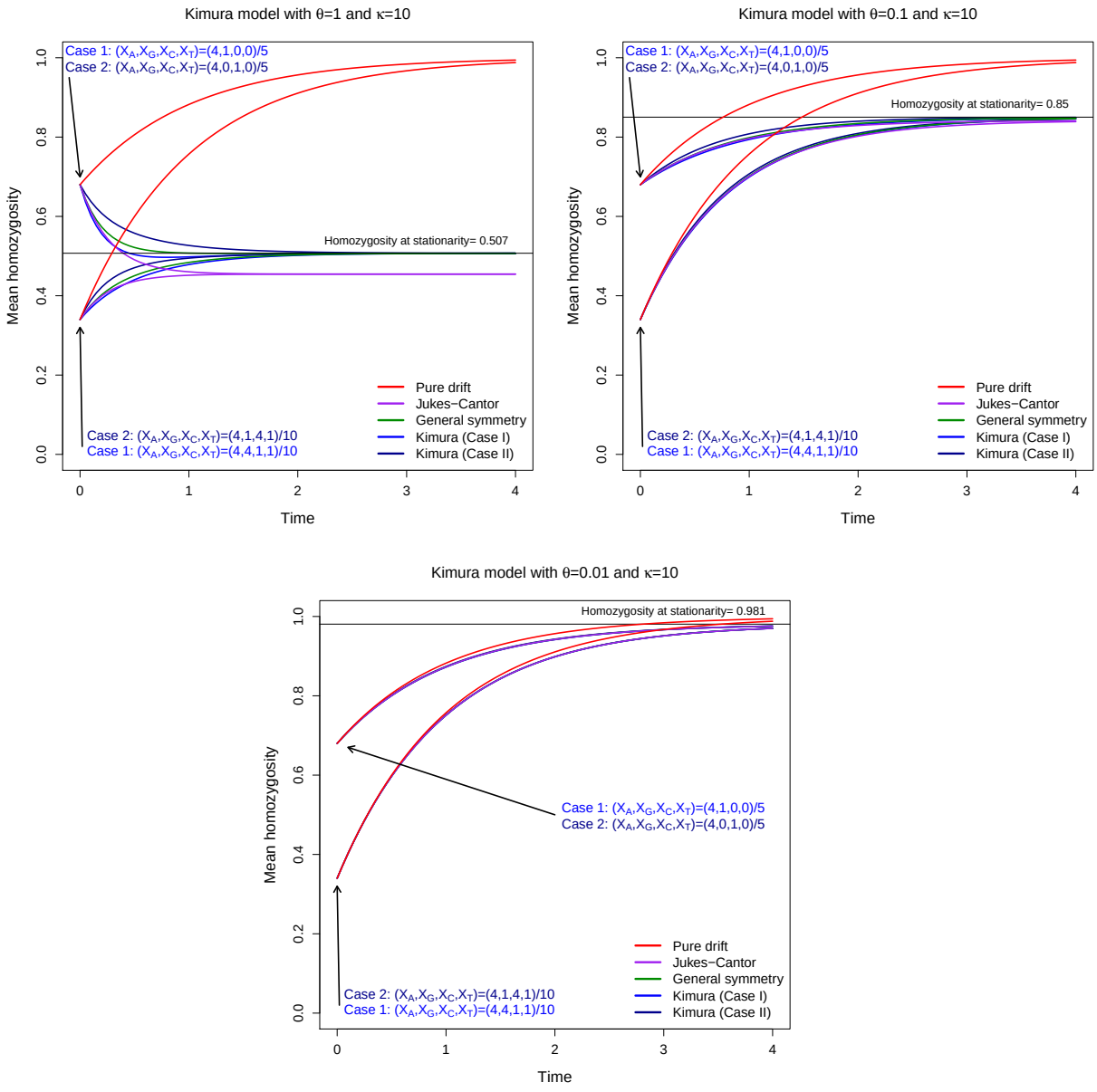

Figure 2: The Kimura model illustrates the various results on the expected homozygosity as a function of time. The blue curves give the expected homozygosity based on (12) and our main Theorem 1 for the Kimura model (4) with $\kappa = 10$ and three different values of $\theta = 1$, $\theta = 0.1$, and $\theta = 0.01$. The green curve is the expected homozygosity with permutation of the initial frequencies (Theorem 9), and is a linear combination of the two corresponding blue curves (see text after Theorem 9). The Jukes-Cantor model (in purple), the pure drift model (in red) and the homozygosity at stationarity (black horizontal lines) are based on Corollary 8, Corollary 5 and Theorem 7.

Two main forces are driving the allele frequency distribution in the Wright-Fisher model with mutation, namely genetic drift and introduction of new mutations. With a small mutation rate genetic drift is the most important force, and the actual mutational process is less important, as can be seen from the plot with the mutation rate $\theta = 0.01$. In this case the expected homozygosity is similar in all situations. For a large mutation rate, however, taking the mutational process into account is crucial, and biases in the mutation process (e.g. transitions versus transversions) should also be accounted for. This is evident from the plots of the Kimura model with $\theta = 1$ and $\kappa = 10$.

We begin our analyses of homozygosity with a small observation.

Corollary 5. Expected homozygosity in the pure drift model
In the pure drift model the homozygosity is

$$\mathrm{E}[F(t)|F(0)] = 1 - e^{-t}(F(0) - 1).$$

Proof. The result follows easily by inserting the formulas from Corollary 3 in (12). □

Griffiths (1980) considers general symmetric mutation models where $U = U'$. For general symmetric mutation models the following Lemma gives the expected homozygosity.

Lemma 6. Expected homozygosity in the general symmetric mutation model
In the general symmetric mutation model ($U = U'$) the homozygosity is given by

$$\mathrm{E}[F(t)|x(0)] = \int_0^t e^{-s}\mathrm{trace}\Big(\mathrm{diag}\{x(0)e^{Q(t-s)}\}(e^{2Qs})\Big)ds + e^{-t}x(0)e^{2Qt}x(0)'. \qquad (13)$$

Proof. We need to insert the result from Theorem 1 in (12). Since Q is symmetric $(e^{Qs})' = e^{Qs}$. Recall that generally $\mathrm{trace}(AB) = \mathrm{trace}(BA)$. Furthermore the integral in (3) is originally derived from a sum and since $\mathrm{trace}(A + B) = \mathrm{trace}(A) + \mathrm{trace}(B)$ we can interchange the trace and the integral. We thus have

$$\mathrm{trace}\Big(\int_0^t e^{-s}(e^{Qs})'\mathrm{diag}\{x(0)e^{Q(t-s)}\}(e^{Qs})ds\Big) = \int_0^t e^{-s}\mathrm{trace}\Big(\mathrm{diag}\{x(0)e^{Q(t-s)}\}(e^{2Qs})\Big)ds,$$

and

$$\mathrm{trace}\Big((e^{Qt})'x(0)'x(0)e^{Qt}(1 - e^{-t})\Big) = (1 - e^{-t})x(0)e^{2Qt}x(0)',$$

and the result is established. □

The first result from Griffiths (1980) is his Theorem 1:

Theorem 7. Expected homozygosity in the stationary phase for the general symmetric mutation model (Theorem 1 in Griffiths (1980))

In the stationary phase of the general symmetric mutation model ($U = U'$) the expected homozygosity is given by

$$\mathrm{E}[F] = \frac{1}{K}\Big\{1 + \sum_{j=1}^{K-1} \frac{1}{1-2\lambda_j}\Big\}$$

where $\lambda_1, \ldots, \lambda_{K-1}, 0$ are the eigenvalues of $Q = N(U - I)$. We note that the expected homozygosity at stationarity is the average of a simple function of the eigenvalues of the rate matrix.

Proof. Recall that for a symmetric probability matrix the stationary distribution is $(1, \ldots, 1)/K = \mathbf{e}/K$ and $(\mathbf{e}/K)e^{Qs} = \mathbf{e}/K$. Let $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_{K-1}, 0)$ be the vector of eigenvalues (hence all entries are smaller or equal to zero) and V the matrix of corresponding eigenvectors such that $Q = V\mathrm{diag}(\lambda)V'$ and $e^{Qs} = V\mathrm{diag}(e^{\lambda s})V'$ where $VV' = I$.

We get

$$\begin{aligned}
\mathrm{E}[F(t)|x(0) = \mathbf{e}/K] &\overset{(\star)}{=} \int_0^t e^{-s}\mathrm{trace}\Big(\mathrm{diag}\{\mathbf{e}/K\}(e^{2Qs})\Big)ds + e^{-t}(\mathbf{e}/K)(\mathbf{e}/K)' \\
&= \frac{1}{K}\int_0^t e^{-s}\mathrm{trace}(e^{2Qs})ds + e^{-t}/K \\
&= \frac{1}{K}\int_0^t e^{-s}\mathrm{trace}\Big(\mathrm{diag}(e^{2\lambda s})\Big)ds + e^{-t}/K \\
&= \frac{1}{K}\sum_{j=1}^{K}\frac{1-e^{-(1-2\lambda_j)t}}{1-2\lambda_j} + e^{-t}/K \\
&= \frac{1}{K}\Big\{1 - e^{-t} + \sum_{j=1}^{K-1}\frac{1-e^{-(1-2\lambda_j)t}}{1-2\lambda_j}\Big\} + e^{-t}/K \\
&= \frac{1}{K}\Big\{1 + \sum_{j=1}^{K-1}\frac{1-e^{-(1-2\lambda_j)t}}{1-2\lambda_j}\Big\} \\
&\underset{t\to\infty}{\longrightarrow} \frac{1}{K}\Big\{1 + \sum_{j=1}^{K-1}\frac{1}{1-2\lambda_j}\Big\},
\end{aligned}$$

where in $(\star)$ we used Lemma 6. □

The second result in Griffiths (1980) is concerned with the transient phase and the Jukes-Cantor mutation model.

Corollary 8. Expected homozygosity in the transient phase for the Jukes-Cantor mutation model (Corollary 3 in Griffiths (1980))

In the transient phase of the Jukes-Cantor mutation model ($U_{ij} = u/(K-1)$ *for* $i \neq j$) *the expected homozygosity is determined by*

$$\mathrm{E}[F(t)|x(0)] = K^{-1} + K^{-1}(K-1)(1+\epsilon)^{-1}\Big(1 - e^{-\tau(1+\epsilon)}\Big) + (F(0) - K^{-1})e^{-\tau(1+\epsilon)},$$

where $\epsilon = 2NuK/(K-1)$ *and* $F(0) = \sum_{k=1}^{K} x_k(0)^2$.

Proof. The proof is another application of Lemma 6 and can be found in Appendix E. □

The third result in Griffiths (1980) is also concerned with the transient phase but for the general symmetric model. The result requires some notation. Let $\sigma_k, k = 1, \ldots, K!$, denote the $K!$ permutation matrices that permute the entries in a vector. For example if $K = 3$ then the possible six permutations of (1,2,3) are (1,2,3), (1,3,2), (2,1,3), (2,3,1), (3,1,2), and (3,2,1). The permutation matrix that transforms (1,2,3) to (1,2,3) is the 3×3 identity matrix $\sigma_1 = I$, and the permutation matrix that transforms (1,2,3) to (1,3,2) is

$$\sigma_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

We can now formulate the theorem.

Theorem 9. Expected homozygosity in the transient phase for the general symmetric mutation model with permutation (Theorem 2 in Griffiths (1980)) *In the general symmetric mutation model* ($U = U'$) *with permutation of the initial frequencies the expected homozygosity is given by*

$$(K!)^{-1}\sum_{k=1}^{K!}\mathrm{E}[F(t)|x(0)\sigma_k] = \frac{1}{K}\Big\{1 + \sum_{i=1}^{K-1}\frac{(1 - e^{-t(1-2\lambda_i)})}{(1-2\lambda_i)}\Big\} + \frac{(F(0) - 1/K)}{K-1}\sum_{i=1}^{K-1} e^{-t(1-2\lambda_i)}, \quad (14)$$

where $\lambda_1, \ldots, \lambda_{K-1}$ *are the eigenvalues of* $Q = N(U - I)$ *and* $F(0)$ *is the initial homozygosity.*

Proof. The proof is yet another application of Lemma 6 and can be found in Appendix F. □

Now consider again Figure 2 and note that because of the symmetry in the Kimura model the expected homozygosity for initial frequency (4,1,0,0)/5 is the same as the expected homozygosity for initial frequencies (1,4,0,0)/5, (0,0,4,1)/5 and (0,0,1,4)/5. Similarly the expected homozygosity is the same for initial frequency (4,0,1,0)/5 and initial frequencies (0,4,1,0)/5, (0,4,0,1)/5, (4,0,0,1)/5, (1,0,4,0)/5, (0,1,4,0)/5, (1,0,0,4)/5 and (0,1,0,4)/5. Thus Griffiths' result in Theorem 9 on the permuted initial frequencies (shown in green) is a weighted sum of the expected homozygosity with initial frequency (4,1,0,0)/5 and with initial frequency (4,0,1,0)/5, where the weights are 1/3 and 2/3, respectively. In Figure 2 we also show the situation where the initial frequency is (4,4,1,1)/10. Again Griffiths' result for the permuted initial frequencies (shown in green) is a weighted sum of (4,4,1,1) with weight 1/3 and (4,1,4,1)/10 with weight 2/3.

4 The hierarchical Beta model

In Section 2 we showed that the Dirichlet model is not appropriate for the distribution of allele frequencies in the Wright-Fisher model with Kimura mutations. In this section we introduce the hierarchical Beta model. The model is adequate for loci with four alleles and by construction is expected to fit the Wright-Fisher with Kimura mutation. We show that the model is analytically tractable and describe in detail an application to estimation of the scaled number of generations.

4.1 Definition and basic properties

Consider a locus with four alleles with frequencies $x = (x_1, x_2, x_3, x_4)$ corresponding to `A,G,C,T`. Define the independent beta-distributed stochastic variables ω, η_1 and η_2 according to

$$\begin{aligned} \omega &\sim \text{Beta}(\phi_a\mu_a, \phi_a(1-\mu_a)) \\ \eta_1 &\sim \text{Beta}(\phi_1\mu_1, \phi_1(1-\mu_1)) \\ \eta_2 &\sim \text{Beta}(\phi_2\mu_2, \phi_2(1-\mu_2)), \end{aligned}$$

with mean, variance and interpretation given in Table 1, and define the joint distribution of x hierarchically according to Table 2.

variable	mean	variance	interpretation
ω	μ_a	$V_a = \mu_a(1-\mu_a)/(\phi_a+1)$	purine (`A+G`) fraction
η_1	μ_1	$V_1 = \mu_1(1-\mu_1)/(\phi_1+1)$	`A` fraction of purines
η_2	μ_2	$V_2 = \mu_2(1-\mu_2)/(\phi_2+1)$	`C` fraction of pyrimidines (`C+T`)

Table 1: Variables, corresponding parameters and their interpretation in the hierarchical Beta model.

allele frequency	definition	mean	variance
x_1	$\omega\eta_1$	$\mu_a\mu_1$	$\mu_1^2V_a + \mu_a^2V_1 + V_aV_1$
x_2	$\omega(1-\eta_1)$	$\mu_a(1-\mu_1)$	$(1-\mu_1)^2V_a + \mu_a^2V_1 + V_aV_1$
x_3	$(1-\omega)\eta_2$	$(1-\mu_a)\mu_2$	$\mu_2^2V_a + (1-\mu_a)^2V_2 + V_aV_2$
x_4	$(1-\omega)(1-\eta_2)$	$(1-\mu_a)(1-\mu_2)$	$(1-\mu_2)^2V_a + (1-\mu_a)^2V_2 + V_aV_2$

Table 2: Definition, mean and variance of nucleotide frequencies in the hierarchical Beta model.

In Table 2 the mean is a result of independence between ω, η_1 and η_2. The variance

of x_1 is calculated from

$$\begin{aligned}\mathrm{Var}(x_1) &= \mathrm{Var}(\omega\eta_1) = \mathrm{Var}[\mathrm{E}(\omega\eta_1|\omega)] + \mathrm{E}[\mathrm{Var}(\omega\eta_1|\omega)] \\ &= \mathrm{Var}[\omega\mu_1] + \mathrm{E}[\omega^2 V_1] \\ &= \mu_1^2 V_a + \mu_a^2 V_1 + V_a V_1,\end{aligned}$$

and similar calculations apply for the other three variances.

The covariance between x_1 and x_2 is given by

$$\begin{aligned}\mathrm{Cov}(x_1, x_2) &= \mathrm{Cov}[\mathrm{E}(\omega\eta_1|\omega), \mathrm{E}(\omega(1-\eta_1)|\omega)] + \mathrm{E}[\mathrm{Cov}(\omega\eta_1, \omega(1-\eta_1)|\omega)] \\ &= \mathrm{Cov}(\omega\mu_1, \omega(1-\mu_1)) - \mathrm{E}[\omega^2 \mathrm{Var}(\eta_1)] \\ &= \mu_1(1-\mu_1)V_a - V_a V_1 - \mu_a^2 V_1.\end{aligned}$$

We note that the covariance between x_1 and x_2 can be positive; recall that positive covariances are not possible in the Dirichlet distribution.

The covariance between x_1 and x_3 is

$$\begin{aligned}\mathrm{Cov}(x_1, x_3) &= \mathrm{Cov}[\mathrm{E}(\omega\eta_1|\omega), \mathrm{E}((1-\omega)\eta_2)|\omega)] + \mathrm{E}[\mathrm{Cov}(\omega\eta_1, (1-\omega)\eta_2|\omega)] \\ &= \mathrm{Cov}(\omega\mu_1, (1-\omega)\mu_2) - \mathrm{E}[\omega(1-\omega)\mathrm{Cov}(\eta_1, \eta_2)] \\ &= -\mu_1\mu_2 V_a.\end{aligned}$$

In Table 3 we summarize the covariance structure of the model.

variables	covariance between variables
x_1, x_2	$\mu_1(1-\mu_1)V_a - V_a V_1 - \mu_a^2 V_1$
x_1, x_3	$-\mu_1\mu_2 V_a$
x_1, x_4	$-\mu_1(1-\mu_2)V_a$
x_2, x_3	$-(1-\mu_1)\mu_2 V_a$
x_2, x_4	$-(1-\mu_1)(1-\mu_2)V_a$
x_3, x_4	$\mu_2(1-\mu_2)V_a - V_a V_2 - (1-\mu_a)^2 V_2$

Table 3: Covariance between nucleotide frequencies in the hierarchical Beta model.

4.2 Moment-based parameter estimation

Having calculated the means and (co)variances of the model we now discuss how the parameters are determined. We first note that the mean of (x_1, x_2, x_3, x_4) is a bijective function of (μ_a, μ_1, μ_2). The means (μ_a, μ_1, μ_2) are therefore completely determined.

The remaining parameters V_a, V_1 and V_2 are determined as follows. We determine V_a from

$$\mathrm{Var}(x_1 + x_2) = V_a, \tag{15}$$

we determine V_1 from

$$\text{Var}(x_1) + \text{Var}(x_2) = [\mu_1^2 + (1-\mu_1)^2]V_a + 2\big[\mu_a^2 + V_a\big]V_1, \tag{16}$$

and we determine V_2 from

$$\text{Var}(x_3) + \text{Var}(x_4) = [\mu_2^2 + (1-\mu_2)^2]V_a + 2\big[(1-\mu_a)^2 + V_a\big]V_2. \tag{17}$$

We now use Table 2 and Table 3 to determine the variance-covariance structure of x.

4.3 Likelihood for a sample of allele counts

In reality we do not observe the allele frequencies (x_1, x_2, x_3, x_4), but a sample of allele counts $c = (c_1, c_2, c_3, c_4)$. Let $\xi = (\alpha_a, \beta_a, \alpha_1, \beta_1, \alpha_2, \beta_2)$ denote the parameters in the hierarchical Beta model. In this section we determine the likelihood $L(c|\xi)$ for a sample c from the hierarchical Beta model conditional on the parameters ξ. Instead of parameterising the hierarchical Beta model in terms of the means and (scaled) variances $(\mu_a, \phi_a, \mu_1, \phi_1, \mu_2, \phi_2)$ we use the shape parameters $\xi = (\alpha_a, \beta_a, \alpha_1, \beta_1, \alpha_2, \beta_2)$ of the Beta distribution such that e.g. $\mu_a = \alpha_a/(\alpha_a + \beta_a)$ and $\phi_a = \alpha_a + \beta_a$.

Let $p(x|\xi)$ be the density of the allele frequencies $x = (x_1, x_2, x_3, x_4)$ conditional on the parameters ξ, and let $L(c|x)$ be the likelihood of the sample c conditional on the frequencies x. The sample likelihood $L(c|\xi)$ is given by

$$L(c|\xi) = \int_x L(c|x)p(x|\xi)dx.$$

In the subsections below we determine $L(c|x)$, $p(x|\xi)$, and finally $L(c|\xi)$. In particular we show that in the hierarchical Beta model the sample likelihood is a product of three Beta-binomial distributions.

4.3.1 Likelihood $L(c|x)$ for c conditional on x

We observe the allele counts $c = (c_1, c_2, c_3, c_4)$ where $c_1 + c_2 + c_3 + c_4 = n$. The likelihood for a sample conditional on the underlying frequency vector $x = (x_1, x_2, x_3, x_4)$ is given by the multinomial distribution $c \sim \text{Mult}(n, x)$. An equivalent description is that the counts follow the distributions

$$\begin{aligned} c_1 + c_2 &\sim \text{Bin}(n, x_1 + x_2) \\ c_1|(c_1 + c_2) &\sim \text{Bin}(c_1 + c_2, x_1/(x_1 + x_2)) \\ c_3|(c_1 + c_2) &\sim \text{Bin}(n - c_1 - c_2, x_3/(1 - x_1 - x_2)), \end{aligned} \tag{18}$$

and using this description the natural summary of the data is in terms of the vector $(c_1 + c_2, c_1, c_3)$, and we have

$$L(c_1 + c_2, c_1, c_3|x) = P(c_1 + c_2|x_1 + x_2)P(c_1|c_1 + c_2, x_1 + x_2, x_1)P(c_3|c_1 + c_2, x_1 + x_2, x_3). \tag{19}$$

4.3.2 Density $p(x|\xi)$ for x conditional on ξ

The density of (x_1, x_2, x_3) can be found by taking advantage of the conditional properties of the hierarchical Beta model. By noting that

$$\begin{aligned} x_1 + x_2 &\sim \text{Beta}(\alpha_a, \beta_a) \\ x_1|(x_1 + x_2) &\sim (x_1 + x_2)\text{Beta}(\alpha_1, \beta_1) \\ x_3|(x_1 + x_2) &\sim (1 - x_1 - x_2)\text{Beta}(\alpha_2, \beta_2) \end{aligned} \tag{20}$$

we get the joint density for $(x_1 + x_2, x_1, x_3)$ from

$$p(x_1 + x_2, x_1, x_3) = p(x_1 + x_2)p(x_1|x_1 + x_2)p(x_3|x_1 + x_2). \tag{21}$$

We note that this formula also holds for the joint density of (x_1, x_2, x_3) because the absolute value of the determinant of the Jacobian $|J_g(x_1, x_2, x_3)|$ for the transformation

$$g(x_1, x_2, x_3) = (x_1 + x_2, x_1, x_3)$$

equals one. In conclusion the density of (x_1, x_2, x_3) is given by

$$\begin{aligned} &p(x_1, x_2, x_3) = \\ &f_{\text{B}(\alpha_a,\beta_a)}(x_1 + x_2)\frac{1}{x_1 + x_2} f_{\text{B}(\alpha_1,\beta_1)}\big(\frac{x_1}{x_1 + x_2}\big)\frac{1}{1 - x_1 - x_2} f_{\text{B}(\alpha_2,\beta_2)}\big(\frac{x_3}{1 - x_1 - x_2}\big) \end{aligned} \tag{22}$$

where e.g. $f_{\text{B}(\alpha_a,\beta_a)}(\cdot)$ is the density function for the Beta-distribution with shape parameters (α_a, β_a).

4.3.3 Sample likelihood $L(c|\xi) = \int_x L(c|x)p(x|\xi)dx$.

Now consider the likelihood

$$L(c_1 + c_2, c_1, c_3|\xi) = \int_x L(c_1 + c_2, c_1, c_3|x_1 + x_2, x_1, x_3)p(x_1 + x_2, x_1, x_3|\xi)dx \tag{23}$$

where $L(c_1 + c_2, c_1, c_3|x_1 + x_2, x_1, x_3)$ is given by (18) and (19), and $p(x_1 + x_2, x_1, x_3|\xi)$ is given by (20) and (21). The likelihood becomes a product of three beta-binomial distributions

$$L(c_1 + c_2, c_1, c_3) = f_{\text{BB}(n,\alpha_a,\beta_a)}(c_1 + c_2) f_{\text{BB}(c_1+c_2,\alpha_1,\beta_1)}(c_1) f_{\text{BB}(c_3+c_4,\alpha_2,\beta_2)}(c_3), \tag{24}$$

where e.g. $f_{\text{BB}(n,\alpha_a,\beta_a)}(k)$ is the probability function of the Beta-binomial distribution with n trials and shape parameters α_a and β_a.

4.3.4 Stationary distribution for symmetric mutation model

As an illustration of our moment-based parameter estimation and corresponding likelihood function we consider the stationary distribution for the symmetric mutation model. In the special case of a Jukes-Cantor mutation model, the stationary distribution for (x_1, x_2, x_3, x_4) is a Dirichlet with parameter vector $(\alpha, \alpha, \alpha, \alpha)$, and the marginal distributions are Beta with shape parameters $(\alpha, 3\alpha)$ (e.g. Ewens, 2004, p. 194-195).

We get $\mu_a = \mu_1 = \mu_2 = 1/2$. Equation (15) becomes

$$V_a = \mathrm{Var}(x_1 + x_2) = \mathrm{Var}(x_1) + \mathrm{Var}(x_2) + 2\mathrm{Cov}(x_1, x_2) = \frac{1}{4(4\alpha + 1)},$$

and therefore $\phi_a = 4\alpha$. Equation (16) becomes

$$\frac{1}{2}V_a + 2\big(\frac{1}{4} + V_a\big)V_1 = \mathrm{Var}(x_1) + \mathrm{Var}(x_2) = \frac{3/2}{4(4\alpha + 1)} = \frac{3}{2}V_a,$$

which implies

$$V_1 = \frac{V_a}{2(\frac{1}{4} + V_a)} = \frac{1}{2(4\alpha + 2)} = \frac{1}{4(2\alpha + 1)},$$

and therefore $\phi_1 = 2\alpha$, and finally $\phi_1 = \phi_2$ by symmetry (or using equation (17)).

Now consider the joint distribution of (x_1, x_2, x_3, x_4) in the hierarchical Beta model with parameters

$$(\mu_a, \phi_a, \mu_1, \phi_1, \mu_2, \phi_2) = (1/2, 4\alpha, 1/2, 2\alpha, 1/2, 2\alpha).$$

Inserting in (22) we get a Dirichlet distribution with parameters $(\alpha, \alpha, \alpha, \alpha)$, and we conclude that the hierarchical Beta approximation is exact in this particular case.

4.4 Application: Estimation of scaled number of generations

To illustrate the hierarchical Beta model we consider the problem of estimating the scaled number of generations (time) in the Wright-Fisher with Kimura mutation process. We assume the initial frequency $x(0) = (16, 2, 1, 1)/20$ is known. Similarly, the mutation rate $\theta = \alpha + 2\beta = 1$ and the transition-transversion rate ratio $\kappa = \alpha/\beta = 10$ are also assumed known. We then simulate a total of 500 independent loci for a scaled number of generations $t = 0.4$ (in the simulation the population size is $N = 1000$). The full data thus consists of 500 frequency vectors of length four corresponding to the frequency of (A,G,C,T). We also sampled count data from the full data. The count data consist of 60 or 20 samples from the multinomial distribution with the simulated frequency vector from each of the 500 loci.

We consider the likelihood as a function of time when the initial frequency and mutation model is fixed. The solid curve in the left figure in Figure 3 shows the log-likelihood (22) based on the fully observed allele frequencies. The dashed and dotted

curves are the log-likelihood (24) based on samples of size 60 or 20. We note that the likelihood curves based on more detailed information are more peaked, but the maximum likelihood value is stable and close to the true value.

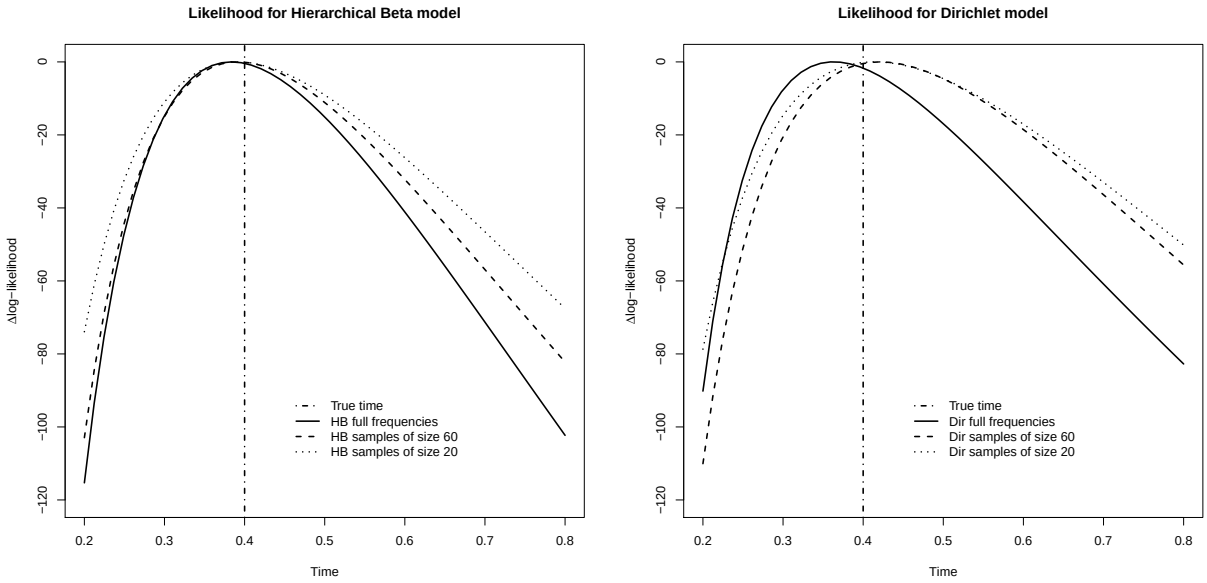

Figure 3: Likelihood function for the scaled number of generations (time) in the Wright-Fisher with Kimura mutation process. Left: Hierarchical Beta model. Right: Dirichlet model.

The right plot in Figure 3 shows the likelihood for the Dirichlet model. The solid curve is the log-likelihood based on the fully observed allele frequencies, and the dashed and dotted curves are based on the two samples. The four Dirichlet parameters are determined in a similar fashion as for the hierarchical Beta model. In particular the means determine three of the parameters, and the last parameter (the concentration parameter) is determined from the variances of the Wright-Fisher with Kimura mutation process. The likelihood for a sample from the Dirichlet model is determined by the Dirichlet-multinomial model; see Gaggiotti and Foll (2010) for more information about parameter estimation in this model. We observe that the maximum likelihood time estimates based on the Dirichlet models are more variable than for the hierarchical Beta model, but they do give reasonable values.

In Figure 4 we show the empirical and marginal distributions from the hierarchical Beta and Dirichlet models. The marginal distributions from the hierarchical Beta and Dirichlet distributions are very similar.

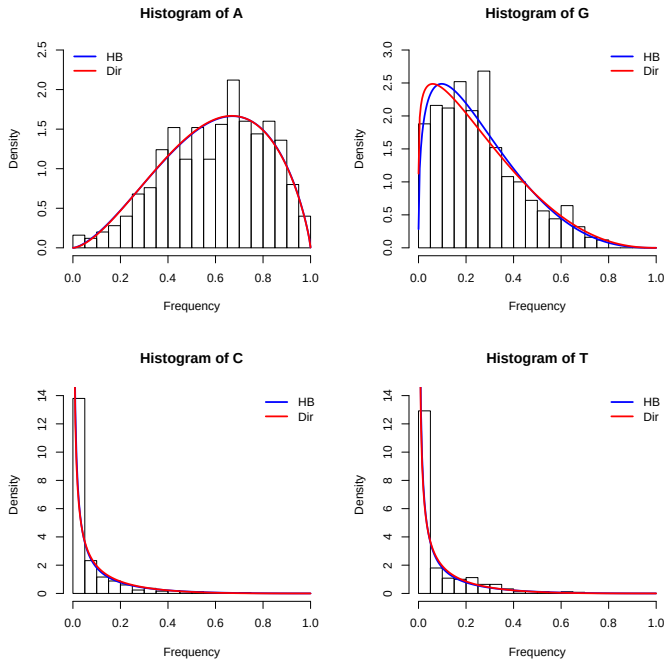

Figure 4: Marginal distributions for the Wright-Fisher with Kimura mutation model. The lines are the marginal densities for the hierarchical Beta and Dirichlet approximations.

5 Simulation studies

5.1 Transition density

Matching the moments does not guarantee that the hierarchical Beta is a good approximation of the transition density of the Wright-Fisher model with Kimura mutations. To get insight into the quality of the approximation in different situations, we simulated both the Wright-Fisher model and its hierarchical Beta approximation and compared the generated samples. Additionally, we also simulated samples from the Dirichlet approximation to see how large effect the correct covariances of the hierarchical Beta have on the accuracy.

We simulated the Wright-Fisher model with Kimura mutations for $2N$ generations for each of the 108 different combinations of parameters as indicated in Table 4. For each parameter combination we generated 10^6 replicates and recorded the allele frequencies at 15 different times. We generated 10^6 samples from each corresponding Dirichlet and hierarchical Beta approximation.

Evaluation of the quality of the approximations was carried out by comparing the empirical cumulative distribution functions (cdf) computed from the samples. A total of 2000 random points were sampled from a uniform distribution on the 3-dimensional

Parameter	N	κ	$\theta = N(2+\kappa)u$	$x(0)$
Value	60	1	1	$(0.25, 0.25, 0.25, 0.25)$
	240	2	0.1	$(0.8, 0.2, 0, 0)$
	960	10	0.01	$(0.4, 0.1, 0.4, 0.1)$
				$(0.4, 0.4, 0.1, 0.1)$

Table 4: Parameter values used in the simulation study of the transition density.

simplex and the empirical cdf was evaluated at these points. The root mean squared difference (RMSD) between the two empirical cdfs was calculated for each comparison.

Figure 5 summarizes the differences between the approximations and the Wright-Fisher model in several cases. With a high mutation rate $\theta = 1$ and $\kappa > 1$, the hierarchical Beta clearly outperforms the Dirichlet. When the mutation rate is $\theta = 0.1$ and $\kappa > 1$ the two approximations behave similarly for short time scales, but with longer times the hierarchical Beta is more accurate. The difference between the two approximations is smaller the closer the initial frequency is to uniform. Interestingly, with $x(0) = (0.8, 0.2, 0, 0)$ the hierarchical Beta was more accurate than the Dirichlet even when mutation rate was low ($\theta < 0.1$) and $\kappa = 1$. This finding is somewhat surprising because the covariance structure is symmetric in this case. The RMSD between the true distribution and the hierarchical Beta approximations peaked with a scaled time around 0.5 or 1 depending on the initial frequency $x(0)$. Full results of all the simulations are shown in the Supplementary Information, with the exception of the case $N = 240$, which produced almost identical results as $N = 960$.

5.2 Stationary distribution

We investigated the accuracy of the approximations in the stationary phase of the Wright-Fisher model with Kimura mutations by using a similar simulation study as with the transient phase. We fixed $N = 960$ and used the parameter values for κ and θ shown in Table 4. For each parameter combination we simulated 10^6 replicates from the Wright-Fisher model in the stationary phase by first drawing initial allele frequencies from the Dirichlet approximation and then simulating the Wright-Fisher model for $10N$ generations. The allele frequencies were recorded at $5N$ and $10N$ generations. We sampled the initial frequencies from the Dirichlet instead of the hierarchical Beta approximation to reduce the possibility that the better performance of hierarchical Beta was caused by the initial frequencies. In each case we simulated 10^7 samples from both the Dirichlet and the hierarchical Beta approximations to the stationary distribution of the Wright-Fisher model. The samples were compared similarly as in the transient case by evaluating the empirical cumulative distribution functions on 2000 random points and recording the RMSD to the Wright-Fisher model.

The root mean squared difference between the two approximations and the Wright-Fisher model are shown in Figure 6. With $\kappa = 1$ the hierarchical Beta and Dirichlet

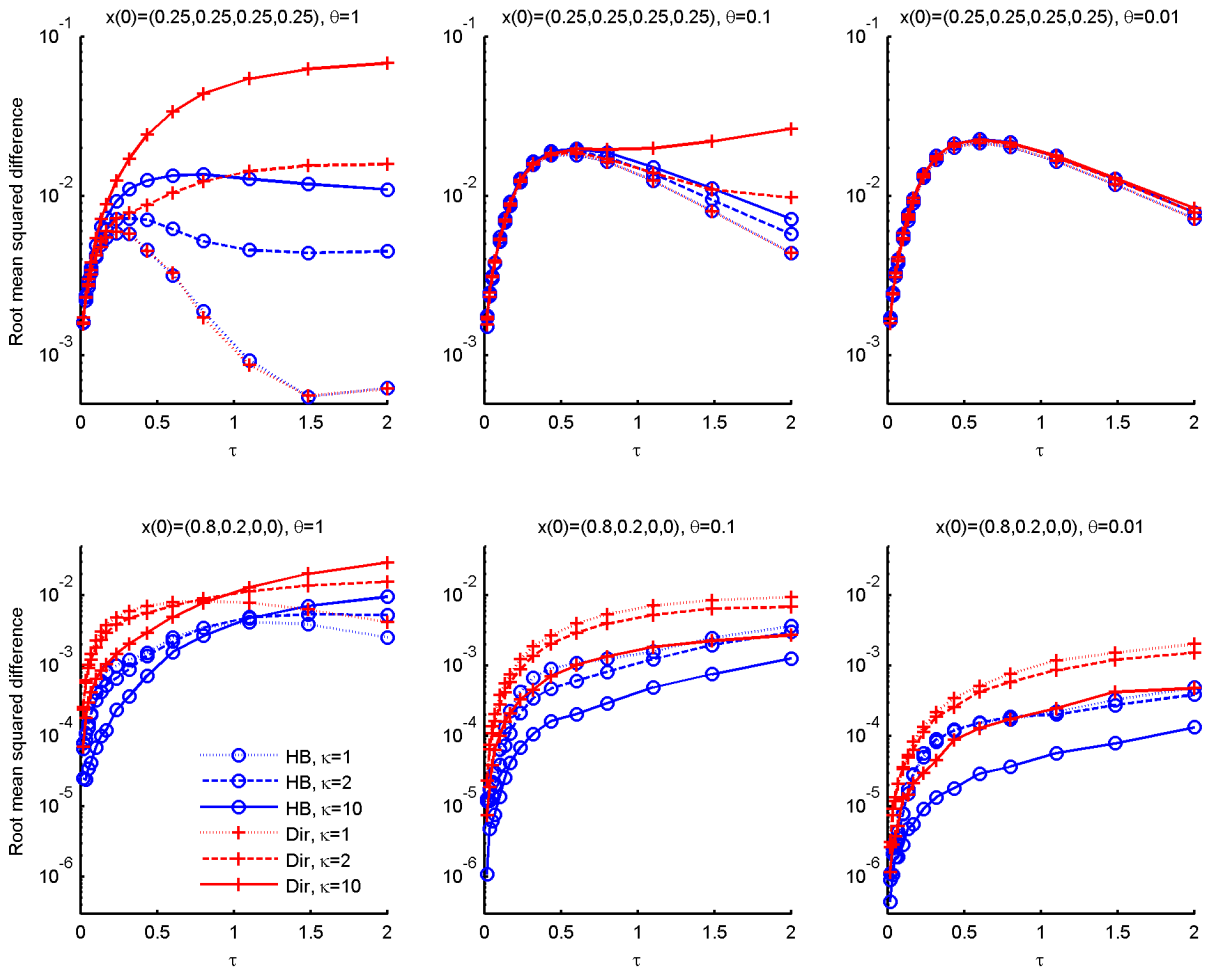

Figure 5: Transient density approximations. The root mean squared difference of the hierarchical Beta (blue circles) and Dirichlet (red crosses) approximations against the Wright-Fisher process with Kimura mutations. Each panel shows the RMSD as a function of time for values of $x(0)$ and θ indicated above the panel. Different linestyles correspond to different values of κ as shown in the legend.

approximations are identical, which is seen in the results. The approximations are also exact for the Wright-Fisher diffusion in that case (Ewens, 2004) and the differences to the discrete Wright-Fisher model come from the discrete support of the distribution and Monte Carlo error. As the value of κ increases the accuracy of the hierarchical Beta is several times larger than that of the Dirichlet, because the latter is not able to capture the covariance structure of the Wright-Fisher model. With $\theta = 0.01$ the RMSD

between the hierarchical Beta and the Wright-Fisher model is smaller than the RMSD between samples from the Wrigh-Fisher model recorded $5N$ generations apart. This is probably due to the simulations from WF-model recorded at $5N$ generations being further from the actual stationary distribution than the same simulations recorded at $10N$ generations.

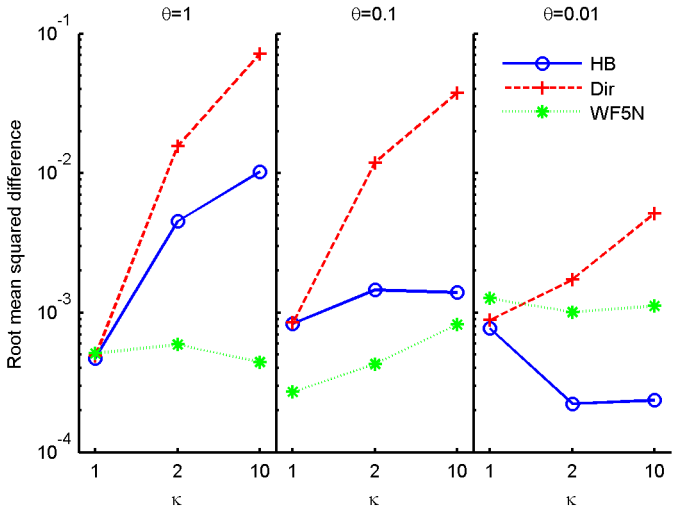

Figure 6: Stationary distribution approximations. The root mean squared difference of the hierarchical Beta (blue circles and solid line) and Dirichlet approximation (red crosses and dashed line) to the stationary distribution of the Wright-Fisher model with $N = 960$ as a function of κ. Green stars and dotted lines lines show the RMSD between realizations of the stationary Wright-Fisher model observed $5N$ generations apart. The value for the mutation parameter θ is indicated above each panel.

6 Conclusion

We have derived a general formulae for the mean and (co)variance of the allele frequencies in the Wright-Fisher with mutation process (Theorem 1), and have considered several special cases of mutation structure, including pure drift (Corollary 3), Jukes-Cantor (Corollary 4) and Kimura (Corollary 2 and Appendix C). We showed that it is generally not possible to fully capture the theoretical covariance structure of the allele frequency distribution in the general multi-allelic case with the Dirichlet distribution, which is the standard approximation to the Wright-Fisher model. This result should be contrasted to the biallelic case, where the Beta distribution is sufficient, and is due to having only one single free parameter for controlling the covariance structure in the Dirichlet distribution. Our results show that in spite of the theoretical mis-match the Dirichlet distribution is generally appropriate to describe the allele frequency distribution in the

case of pure drift, or for the Jukes-Cantor model with either a small mutation rate or a short evolutionary distance. However, for more general situations, the Dirichlet is not flexible enough to capture the (co)variance structure.

We introduced a new statistical approximation for the Wright-Fisher with Kimura mutations, the hierarchical Beta model, which can be parametrized to have approximately the correct theoretical first two moments of the Wright-Fisher model. With extensive simulation studies we demonstrated that the hierarchical Beta model captures the allele frequency distributation very well.

There are two main reasons why the approximations to the Wright-Fisher model with mutation should be considered. First, if the evolutionary timescale is intermediate, such as when studying several subspecies or closely related species, both mutation and genetic drift play important roles in shaping the genetic variation. It is inadequate to model only one of these forces, by either standard phylogenetic methods or by pure drift-based approximations to the Wright-Fisher model. Second, many of the current methods that include both mutation and drift are limited by the number of samples that can be simultaneously analyzed. These include coalescent based methods as well as others such as PoMo (De Maio, Schloetterer and Kosiol, 2013; De Maio, Schrempf and Kosiol, 2015), which extends standard phylogenetic models to account for incomplete lineage sorting and ancestral variation.

The results in this paper were originally motivated by the desire to construct statistical approximations for the transition density of the Wright-Fisher diffusion. However, the moments derived in Theorem 1 also facilitate the study of theoretical properties of the model, as demonstrated in Section 3 by the new proofs and extensions of homozygosity results in Griffiths (1980).

In this work we considered only the first two moments of the transition distribution. Higher order moments could at least in principle be computed in a similar fashion as the first two, although the calculations are more involved (especially in the multi-allelic case). We derived the expression for the mean and variance by repeated use of the formulas for conditional mean and variance. Another route to obtain the moments is to first formulate recursive equations for the moments and second turn the recursions into differential equations by letting the population size N tend to infinity (e.g. Crow and Kimura, 1970, page 336). It could be interesting to investigate this alternative procedure for establishing expressions for the moments. Higher-order moments would be helpful to get insight into the variability in homozygosity. Also, if selection is included in addition to mutation, then the first two moments might not be able to capture the variation in the allele frequencies adequately.

The hierarchical Beta model was introduced as a more accurate approximation to the Wright-Fisher with Kimura mutations than the Dirichlet distribution. If some other mutational model was considered, such as the general time-reversible model (e.g. Felsenstein, 2004, page 204), then the hierarchical Beta model might not be able to capture the (co)variance structure anymore, because it relies on the division of the alleles into two groups (purines and pyrimidines). We emphasize that different approximations need

to be derived for different mutational models, and it remains a challenge to formulate appropriate statistical models for the allele frequency distribution in the Wright-Fisher with mutation models that are more complex than pure drift, Jukes-Cantor or Kimura.

We showed in a simulation study that the hierarchical Beta approximation can be used to estimate the scaled number of generations. A natural continuation of this work would be to implement the hierarchical Beta approximation in a more complex model for inference of population structure.

Acknowledgements

We are grateful to Bob Griffiths for illuminating discussions and careful explanation of details in his 1980 paper, and to two anynomous reviewers who provided constructive suggestions and helpful comments that helped improve the paper. We would also like to thank Freddy Bugge Christiansen, David Balding, Monty Slatkin and Carsten Wiuf for valuable discussions and suggestions. Asger Hobolth is supported by the Danish Research Council (grant number DFF-4002-00382), and Jukka Siren is supported by the Academy of Finland (grant number 273253).

References

Balding, D. and Steele, C. (2015). *Weight-of-evidence for forensic DNA profiling. 2nd edition.* Wiley, Queensland, Australia.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. and RoyChoudhury, A. (2012). Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. *Mol. Biol. Evol.*, **29**(8), 1917–1932.

Cavalli-Sforza, L.L. and Edwards, A.W.F. (1967). Phylogenetic analysis: models and estimation procedures. *Am J Hum Genet*, **19**, 233–257.

Crow, J.F. and Kimura, M. (1970). *An introduction to population genetics theory.* The Blackburn Press, New Jersey, USA.

Degnan, J.H. and Rosenberg, N.A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution*, **24**, 332–340.

De Maio, N., Schlötterer, C. and Kosiol, C. (2013). Linking Great Apes Genome Evolution across Time Scales Using Polymorphism-Aware Phylogenetic Models. *Mol. Biol. Evol.*, **30**, 2249–2262.

De Maio, N., Schrempf, D. and Kosiol, C. (2015). PoMo: An Allele Frequency-based Approach for Species Tree Estimation. *Systematic Biology*, doi: 10.1093/sysbio/syv048.

Edwards, A.W.F and Cavalli-Sforza, L.L. (1964). Reconstruction of evolutionary trees. Phenetic and Phylogenetic Classification. Systematics Association Publ. No. 6, London. Editors: Heywood, V.H. and McNeill, J. pages 67-76.

Ewens, W.J. (2004). *Mathematical Population Genetics.* 2nd Edition. Springer, New York.

Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am J Hum Genet*, **25**, 47.

Felsenstein, J. (2004). *Inferring Phylogenies.* Sinauer Associates, Sunderland, Massachusetts.

Gaggiotti, O.E. and Foll, M. (2010). Quantifying population structure using the F-model. *Molecular Ecology Resources*, **10**, 821–830.

Griffiths, R. (1980). Allele frequencies in multidimensional Wright-Fisher models with a general symmetric mutation structure. *Theoretical Population Biology*, **71**, 51-70.

Gutenkunst, R.N., Hernandez, R.D., Williamson, S. and Bustamante, C.D (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet*, **5**, e1000695.

Heled, J. and Drummond, A.J. (2010). Bayesian Inference of Species Trees from Multilocus Data. *Mol. Biol. Evol.*, **27**, 570-580.

Hobolth, A. and Jensen, J.L. (2005). Statistical inference in evolutionary models of DNA sequences via the EM algorithm. *Statistical applications in Genetics and Molecular Biology*, **4**, 18.

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

Maddison, W.P. (1997). Gene Trees in Species Trees. *Systematic Biology*, **46**, 523-536.

Moler, C and Van Loan, C. (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rewiew*, Vol. 45, No. 1,pp. 3-49.

Nicholson, G., Smith, A.V., Jonsson, F., Gustafsson, K. and Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J. Roy. Stat. Soc.* B, **64**, 695–715.

Pickrell, J. and Pritchard, J.K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet*, **8**, e1002967.

Sirén, J. (2012). Statistical models for inferring the structure and history of populations from genetic data. PhD thesis, University of Helsinki.

Sirén, J., Hanage, W.P. and Corander, J. (2013). Inference on Population Histories by Approximating Infinite Alleles Diffusion. *Mol. Biol. Evol.*, **30**, 457–468.

Sirén, J., Marttinen, P. and Corander, J. (2011). Reconstructing Population Histories from Single Nucleotide Polymorphism Data. *Mol. Biol. Evol.*, **28**, 673–683.

A Variance in the Wright-Fisher model

In this subsection we show the expression (3) for the variance-covariance matrix. Recall from (1) that the dynamics of the number of alleles is given by

$$z(m)|z(m-1) \sim \text{Mult}(N, x(m-1)U),$$

where $x(m) = z(m)/N$ is the allele frequency in generation m. Consider first the variance

$$\begin{aligned} \text{Var}[z_i(m)|z(m-1)] &= N(x(m-1)U)_i(1-(x(m-1)U)_i) \\ &= N(x(m-1)U)_i - N(x(m-1)U)_i(x(m-1)U)_i, \end{aligned}$$

and second the covariance

$$\text{Cov}(z_i(m), z_j(m)|z(m-1)) = -N(x(m-1)U)_i(x(m-1)U)_j = -N(U'x(m-1)'x(m-1)U)_{ij},$$

for $i \neq j$. We thus have

$$\text{Var}[x(m)|x(m-1)]_{ij} = -\frac{1}{N}(x(m-1)U)_i(x(m-1)U)_j + \frac{1}{N}(x(m-1)U)_i 1(i=j),$$

where $1(\cdot)$ is the indicator function. In matrix notation

$$\text{Var}[x(m)|x(m-1)] = -\frac{1}{N}U'(x(m-1)'x(m-1))U + \frac{1}{N}\text{diag}(x(m-1)U).$$

The formulas now become more readable if we write x_m instead of $x(m)$. We use the law of total variance to get

$$\begin{aligned} \text{Var}[x_m] &= \text{E}[\text{Var}(x_m|x_{m-1})] + \text{Var}[\text{E}(x_m|x_{m-1})] \\ &= -\frac{1}{N}U'\text{E}[x'_{m-1}x_{m-1}]U + \frac{1}{N}\text{diag}\{(\text{E}x_{m-1})U\} + U'(\text{Var}x_{m-1})U \\ &= \frac{1}{N}\text{diag}\{(\text{E}x_{m-1})U\} - \frac{1}{N}U'\Big\{\text{Var}x_{m-1} + \text{E}[x'_{m-1}]\text{E}[x_{m-1}]\Big\}U + U'(\text{Var}x_{m-1})U \\ &= \frac{1}{N}\Big[\text{diag}\{(\text{E}x_{m-1})U\} - U'\text{E}[x'_{m-1}]\text{E}[x_{m-1}]U\Big] + \Big(1-\frac{1}{N}\Big)U'(\text{Var}x_{m-1})U. \end{aligned}$$

We now make repeated use of the law of total variance and find

$$\begin{aligned}
\mathrm{Var}[x_m] &= \frac{1}{N}\Big[\mathrm{diag}\{(\mathrm{E}x_{m-1})U\} - U'\mathrm{E}[x'_{m-1}]\mathrm{E}[x_{m-1}]U\Big] + \Big(1-\frac{1}{N}\Big)U'(\mathrm{Var}x_{m-1})U \\
&= \frac{1}{N}\Big[\mathrm{diag}\{(\mathrm{E}x_{m-1})U\} - U'\mathrm{E}[x'_{m-1}]\mathrm{E}[x_{m-1}]U\Big] \\
&\quad +\Big(1-\frac{1}{N}\Big)U'\frac{1}{N}\Big[\mathrm{diag}\{(\mathrm{E}x_{m-1})U\} - U'\mathrm{E}[x'_{m-1}]\mathrm{E}[x_{m-1}]U\Big]U \\
&\quad +\Big(1-\frac{1}{N}\Big)U'\Big(1-\frac{1}{N}\Big)U'(\mathrm{Var}x_{m-1})UU \\
&= \frac{1}{N}\Big[\mathrm{diag}\{(\mathrm{E}x_{m-1})U\} - U'\mathrm{E}[x'_{m-1}]\mathrm{E}[x_{m-1}]U\Big] \\
&\quad +\frac{1}{N}\Big(1-\frac{1}{N}\Big)U'\Big[\mathrm{diag}\{(\mathrm{E}x_{m-2})U\} - U'\mathrm{E}[x'_{m-2}]\mathrm{E}[x_{m-2}]U\Big]U \\
&\quad +\Big(1-\frac{1}{N}\Big)^2(U')^2(\mathrm{Var}x_{m-2})U^2 \\
&= \cdots \\
&= \sum_{i=0}^{m}\frac{1}{N}\Big(1-\frac{1}{N}\Big)^i(U')^i\Big[\mathrm{diag}\{x_0U^{m-1-i}\}\Big]U^i \\
&\quad -\sum_{i=0}^{m}\frac{1}{N}\Big(1-\frac{1}{N}\Big)^i(U')^m x'_0x_0U^m. \qquad (25)
\end{aligned}$$

Recall that $m = tN$ and $Q = N(U - I)$. We can therefore approximate the variance by

$$\mathrm{Var}[x_t] = \int_0^t e^{-s}(e^{Qs})'\mathrm{diag}\{x_0e^{Q(t-s)}\}(e^{Qs})ds - (e^{Qt})'x'_0x_0e^{Qt}(1-e^{-t}). \qquad (26)$$

We suggest computating the first term using an eigenvalue decomposition of the rate matrix; see the next section.

B Computing the (co)variance matrix

We follow Appendix B in Hobolth and Jensen (2005). Assuming reversibility of the mutation process we have

$$\mathrm{diag}(\pi)Q = Q'\mathrm{diag}(\pi),$$

where π is the stationary distribution. The matrix

$$S = \mathrm{diag}(\pi)^{1/2}Q\mathrm{diag}(\pi)^{-1/2}$$

is symmetric and therefore has a computationally robust eigenvalue decomposition

$$S = V\mathrm{diag}(\lambda)V'.$$

It follows that

$$P(t) = e^{Qt} = \mathrm{diag}(\pi^{-1/2}) V \mathrm{diag}(e^{t\lambda}) V' \mathrm{diag}(\pi^{1/2})$$

so that

$$P_{ab}(t) = [e^{Qt}]_{ab} = \Big(\frac{\pi_b}{\pi_a}\Big)^{1/2} \sum_i V_{ai} V_{bi} e^{\lambda_i t}.$$

We now find the following formula for the first term of the variance

$$\Big[\int_0^t e^{-s}(e^{Qs})^* \mathrm{diag}\{x e^{Q(t-s)}\}(e^{Qs}) ds\Big]_{ij} = \sum_k \int_0^t e^{-s}[(e^{Qs})^*]_{ik} \sum_l x_l [e^{Q(t-s)}]_{lk} [e^{Qs}]_{kj} ds$$

$$= \sum_k \int_0^t e^{-s} \Big(\frac{\pi_i}{\pi_k}\Big)^{1/2} \sum_m V_{km} V_{im} e^{\lambda_m s} \sum_l x_l \Big(\frac{\pi_k}{\pi_l}\Big)^{1/2} \sum_n V_{ln} V_{kn} e^{\lambda_n (t-s)} \Big(\frac{\pi_j}{\pi_k}\Big)^{1/2} \sum_r V_{kr} V_{jr} e^{\lambda_r s} ds$$

$$= \sqrt{\pi_i \pi_j} \sum_m V_{im} \sum_l \frac{1}{\sqrt{\pi}_l} x_l \sum_n V_{ln} \sum_r A_{mnr} V_{jr} \sum_k \frac{1}{\sqrt{\pi}_k} V_{km} V_{kn} V_{kr}, \tag{27}$$

where

$$A_{mnr} = \int_0^t e^{-s + \lambda_m s + \lambda_n (t-s) + \lambda_r s} ds = \begin{cases} t e^{\lambda_n t} & \text{if } \lambda_m + \lambda_r = 1 + \lambda_n \\ \frac{e^{(\lambda_m + \lambda_r - 1)t} - e^{\lambda_n t}}{\lambda_m + \lambda_r - 1 - \lambda_n} & \text{otherwise.} \end{cases}$$

Note that expression (27) suggests a fast recursive procedure for evaluating the covariance matrix.

C Variance in the Kimura mutation model

In this appendix we derive the mean and variance from Theorem 1 for the Kimura model. Recall from equation (5) that we can write

$$e^{Qt} = \sum_{i=1}^{3} a_i e^{-b_i t} A_i$$

with obvious definitions of a_i, b_i and A_i. We immediately get the mean $x(0)e^{Qt}$ in the Kimura mutation model.

Recall the general formula for the variance in the Wright-Fisher with mutation model

$$\mathrm{Var}[x(t)|x(0)] = \int_0^t e^{-s}(e^{Qs})' \mathrm{diag}\{x(0) e^{Q(t-s)}\}(e^{Qs}) ds - (e^{Qt})' x(0)' x(0) e^{Qt}(1 - e^{-t}).$$

The last term is easy, and the first term is calculated from

$$\int_0^t e^{-s}(e^{Qs})'\text{diag}\{x(0)e^{Q(t-s)}\}(e^{Qs})ds =$$
$$\int_0^t e^{-s}\sum_{i=1}^3 a_i e^{-b_i s}A_i\text{diag}\Big\{x(0)\sum_{j=1}^3 a_j e^{-b_j(t-s)}A_j\Big\}\sum_{k=1}^3 a_k e^{-b_k s}A_k ds =$$
$$\sum_{i=1}^3\sum_{j=1}^3\sum_{k=1}^3 a_i a_j a_k g_{ijk}(t)A_i\text{diag}\{x(0)A_j\}A_k$$

where

$$g_{ijk}(t) = \int_0^t e^{-s}e^{-b_i s}e^{-b_j(t-s)}e^{-b_k s}ds = \begin{cases} te^{-b_j t} & \text{if } 1+b_i+b_k = b_j \\ \frac{e^{-b_j t}-e^{-(1+b_i+b_k)t}}{1+b_i+b_k-b_j} & \text{otherwise.} \end{cases}$$

These expressions are straight forward to implement.

D Variance in the transient phase for the Jukes-Cantor model

In matrix notation the transition probability matrix for the Jukes-Cantor model is given by

$$P(t) = e^{Qt} = a(t)(I - E/K) + E/K, \tag{28}$$

where E is the $K \times K$ matrix with 1 in every entry and

$$a(t) = \exp(-qKt/(K-1)) = \exp(-\epsilon t/2).$$

We therefore get the mean

$$\text{E}[x(t)] = x(0)e^{Qt} = x(0)\Big[a(t)(I - E/K) + E/K\Big] = e^{-\epsilon t/2}(x(0) - \mathbf{e}/K) + \mathbf{e}/K,$$

where we have used $x(0)E = e$. We now turn to the variance

$$\text{Var}[x(t)] = \int_0^t e^{-s}(e^{Qs})'\text{diag}\{x(0)e^{Q(t-s)}\}(e^{Qs})ds - (e^{Qt})'x(0)'x(0)e^{Qt}(1-e^{-t}). \tag{29}$$

The last term is determined by

$$\begin{aligned}(e^{Qt})'x(0)'x(0)e^{Qt} &= \Big[a(t)(x(0) - \mathbf{e}/K)' + (\mathbf{e}/K)'\Big]\Big[a(t)(x(0) - \mathbf{e}/K) + \mathbf{e}/K\Big] \\ &= e^{-\epsilon t}(x(0) - \mathbf{e}/K)'(x(0) - \mathbf{e}/K) + e^{-\epsilon t/2}(x(0) - \mathbf{e}/K)'\mathbf{e}/K + e^{-\epsilon t/2}(\mathbf{e}/K)'(x(0) - \mathbf{e}/K) + E/K^2.\end{aligned}$$

The first term in the variance becomes

$$\int_0^t e^{-s}\Big(a(s)(I-E/K)+E/K\Big)\Big(a(t-s)(\mathrm{diag}(x(0))-I/K)+I/K\Big)\Big(a(s)(I-E/K)+E/K\Big)ds =$$
$$\Big[\mathrm{diag}(x(0)-\mathbf{e}/K)-(x(0)-\mathbf{e}/K)'\mathbf{e}/K-(\mathbf{e}/K)'(x(0)-\mathbf{e}/K)\Big]\frac{1}{1+\epsilon/2}e^{-\epsilon t/2}\Big(1-e^{-(1+\epsilon/2)t}\Big)+$$
$$\Big[(x(0)-\mathbf{e}/K)'\mathbf{e}/K+(\mathbf{e}/K)'(x(0)-\mathbf{e}/K)\Big]e^{-\epsilon t/2}\Big(1-e^{-t}\Big)+$$
$$\Big[I-E/K\Big]/K\frac{1}{1+\epsilon}\Big(1-e^{-(1+\epsilon)t}\Big)+$$
$$E/K^2(1-e^{-t}),$$

where we have used

$$\int_0^t e^{-s}a^2(s)a(t-s)ds=\frac{1}{1+\epsilon/2}e^{-\epsilon t/2}\Big(1-e^{-(1+\epsilon/2)t}\Big)$$
$$\int_0^t e^{-s}a(s)a(t-s)ds=e^{-\epsilon t/2}\Big(1-e^{-t}\Big)$$
$$\int_0^t e^{-s}a^2(s)ds=\frac{1}{1+\epsilon}\Big(1-e^{-(1+\epsilon)t}\Big)$$
$$\int_0^t e^{-s}ds=1-e^{-t}.$$

The final expression for the variance is

$$\mathrm{Var}[x(t)]= \tag{30}$$
$$\Big[I-E/K\Big]/K\frac{1}{1+\epsilon}\Big(1-e^{-(1+\epsilon)t}\Big)-$$
$$(x(0)-\mathbf{e}/K)'(x(0)-\mathbf{e}/K)e^{-\epsilon t}(1-e^{-t})+$$
$$\Big[\mathrm{diag}(x(0)-\mathbf{e}/K)-(x(0)-\mathbf{e}/K)'\mathbf{e}/K-(\mathbf{e}/K)'(x(0)-\mathbf{e}/K)\Big]e^{-\epsilon t/2}\frac{1}{1+\epsilon/2}\Big(1-e^{-(1+\epsilon/2)t}\Big).$$

E Homozygosity in the transient phase for the Jukes-Cantor model

Corollary 8 follows after an application of Lemma 6. From equation (10) the matrix exponential is

$$e^{Qt}=\frac{1}{K}\Big(1-e^{-\epsilon t/2}\Big)E+e^{-\epsilon t/2}I$$

where $\epsilon=2q/(K-1)=2Nu/(K-1)$. We get

$$\begin{aligned}\mathrm{diag}\{x(0)e^{Q(t-s)}\} &= \frac{1}{K}\Big(1-e^{-\epsilon(t-s)/2}\Big)\mathrm{diag}\{x(0)E\}+e^{-\epsilon(t-s)/2}\mathrm{diag}\{x(0)\}\\ &= \frac{1}{K}\Big(1-e^{-\epsilon(t-s)/2}\Big)I+e^{-\epsilon(t-s)/2}\mathrm{diag}\{x(0)\},\end{aligned} \tag{31}$$

where we have used

$$\mathrm{diag}\{x(0)E\} = \mathrm{diag}\{e\} = I.$$

We now note that

$$\mathrm{trace}(I) = \mathrm{trace}(E) = K$$

and

$$\mathrm{trace}\Big(\mathrm{diag}\{x(0)\}E\Big) = \mathrm{trace}\Big(\mathrm{diag}\{x(0)\}\Big) = 1,$$

and we find

$$\begin{aligned}
&\mathrm{trace}\Big(\mathrm{diag}\{x(0)e^{Q(t-s)}\}(e^{2Qs})\Big) = \frac{1}{K}\Big(1-e^{-\epsilon(t-s)/2}\Big)\frac{1}{K}\Big(1-e^{-\epsilon s}\Big)K + \\
&\quad \frac{1}{K}\Big(1-e^{-\epsilon(t-s)/2}\Big)e^{-2\epsilon s/2}K + e^{-\epsilon(t-s)/2}\frac{1}{K}\Big(1-e^{-\epsilon s}\Big) + e^{-\epsilon(t-s)/2}e^{-\epsilon s} \\
&= \frac{1}{K} + \big(1-\frac{1}{K}\big)e^{-\epsilon s}.
\end{aligned}$$

We therefore have

$$\begin{aligned}
\int_0^t e^{-s}\mathrm{trace}\Big(\mathrm{diag}\{x(0)e^{Q(t-s)}\}(e^{2Qs})\Big)ds &= \int_0^t e^{-s}\Big[\frac{1}{K} + \big(1-\frac{1}{K}\big)e^{-s\epsilon}\Big]ds \\
&= \frac{1}{K}\big(1-e^{-t}\big) + \frac{K-1}{K}\big(1+\epsilon\big)^{-1}\Big(1-e^{-t(1+\epsilon)}\Big). \qquad (32)
\end{aligned}$$

The second term in (13) is easier to calculate. We get

$$\begin{aligned}
x(0)e^{2Qt}x(0)' &= x(0)\Big[\frac{1}{K}\big(1-e^{-t\epsilon}\big)E + e^{-t\epsilon}I\Big]x(0)' \\
&= \frac{1}{K} - \frac{1}{K}e^{-t\epsilon} + F(0)e^{-t\epsilon} \\
&= \frac{1}{K} + (F(0)-K^{-1})e^{-t\epsilon} \qquad (33)
\end{aligned}$$

where we have used $x(0)Ex(0)' = ex(0)' = 1$ and $F(0) = x(0)x(0)'$. Inserting (32) and (33) in (13) we get

$$\begin{aligned}
\mathrm{E}[F(t)|x(0)] &= \frac{1}{K} + \frac{K-1}{K}\Big(1+\epsilon\Big)^{-1}\Big(1-e^{-t(1+\epsilon)}\Big) + (F(0)-K^{-1})e^{-t(1+\epsilon)} \\
&= K^{-1} + K^{-1}(K-1)(1+\epsilon)^{-1}\Big(1-e^{-t(1+\epsilon)}\Big) + (F(0)-K^{-1})e^{-t(1+\epsilon)}.
\end{aligned}$$

F Homozygosity in the transient phase for the general symmetric model

The proof of Theorem 9 is based on Lemma 6 and a direct calculation. From Lemma 6 we get

$$\begin{aligned}(K!)^{-1}\sum_{k=1}^{K!}\mathrm{E}[F(t)|x(0)\sigma_k] &= (K!)^{-1}\sum_{k=1}^{K!}\int_0^t e^{-s}\mathrm{trace}\Big(\mathrm{diag}\{x(0)\sigma_k e^{Q(t-s)}\}(e^{2Qs})\Big)ds + \\ &\quad (K!)^{-1}\sum_{k=1}^{K!} e^{-t}(x(0)\sigma_k)' e^{2Qt}x(0)\sigma_k. \qquad (34)\end{aligned}$$

We begin by calculating the first term in (34)

$$\begin{aligned}\text{First term in (34)} &= \int_0^t e^{-s}\mathrm{trace}\Big(\mathrm{diag}\big\{\big[(K!)^{-1}\sum_{k=1}^{K!}x(0)\sigma_k\big]e^{Q(t-s)}\big\}(e^{2Qs})\Big)ds \\ &= \int_0^t e^{-s}\mathrm{trace}\Big(\mathrm{diag}\{[\mathbf{e}/K]e^{Q(t-s)}\}(e^{2Qs})\Big)ds \\ &= \int_0^t e^{-s}\mathrm{trace}\Big(\mathrm{diag}\{\mathbf{e}/K\}(e^{2Qs})\Big)ds \\ &= \frac{1}{K}\int_0^t e^{-s}\mathrm{trace}(e^{2Qs})ds \\ &= \frac{1}{K}\int_0^t e^{-s}\mathrm{trace}\Big(\mathrm{diag}\big\{e^{2\lambda s}\big\}\Big)ds \\ &= \frac{1}{K}\sum_{i=1}^{K}\frac{1-e^{-(1-2\lambda_i)t}}{1-2\lambda_i} \\ &= \frac{1}{K}\Big(1-e^{-t}+\sum_{i=1}^{K-1}\frac{1-e^{-(1-2\lambda_i)t}}{1-2\lambda_i}\Big) \\ &= \frac{1}{K}\Big(1+\sum_{i=1}^{K-1}\frac{1-e^{-(1-2\lambda_i)t}}{1-2\lambda_i}\Big)-\frac{e^{-t}}{K}.\end{aligned}$$

The second term in (34) is given by

$$\begin{aligned}
\text{Second term in (34)} &= e^{-t}\frac{1}{K!}\sum_{k=1}^{K!} x(0)\sigma_k e^{2Qt}\sigma_k' x(0)' \\
&= e^{-t}\frac{1}{K!}\sum_{k=1}^{K!} \text{trace}\Big(e^{2Qt}\sigma_k' x(0)' x(0)\sigma_k\Big) \\
&= e^{-t}\text{trace}\Big(e^{2Qt}\frac{1}{K!}\sum_{k=1}^{K!}(x(0)\sigma_k)' x(0)\sigma_k\Big) \\
&\overset{(\star)}{=} e^{-t}\Big[\frac{(F(0)-1/K)}{K-1}\Big(1+\sum_{i=1}^{K-1} e^{2\lambda_i t}\Big)+\frac{K(1-F(0))}{K(K-1)}\Big] \\
&= \frac{(F(0)-1/K)}{K-1}\sum_{i=1}^{K-1} e^{-t(1-2\lambda_i)} + e^{-t}\frac{(F(0)-1/K)}{K-1} + e^{-t}\frac{(1-F(0))}{K-1} \\
&= \frac{(F(0)-1/K)}{K-1}\sum_{i=1}^{K-1} e^{-t(1-2\lambda_i)} + \frac{e^{-t}}{K},
\end{aligned}$$

where in $(\star)$ we have used

$$\begin{aligned}
\frac{1}{K!}\sum_{k=1}^{K!}(x(0)\sigma_k)' x(0)\sigma_k &= \frac{1}{K!}\Big[(K-1)!F(0)I + 2(K-2)!\frac{(1-F(0))}{2}(E-I)\Big] \\
&= \frac{1}{K}F(0)I + \frac{(1-F(0))}{K(K-1)}(E-I) \\
&= \frac{(F(0)-1/K)}{K-1}I + \frac{(1-F(0))}{K(K-1)}E.
\end{aligned}$$

Here the first equation is true because on the diagonal we have $F(0)$ a number $K!/K = (K-1)!$ times in the sum, and the off-diagonal entries are

$$\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} x_i(0)x_j(0) = \frac{1-F(0)}{2}$$

a number $K!/\binom{K}{2} = 2(K-2)!$ times. Finally in order to obtain $(\star)$ we note that

$$\text{trace}\Big(e^{Qt}E\Big) = \text{trace}\Big(e^{Qt}\mathbf{e}'\mathbf{e}\Big) = \text{trace}\Big(\mathbf{e}e^{Qt}\mathbf{e}'\Big) = \text{trace}\Big(\mathbf{e}\mathbf{e}'\Big) = K.$$

The Theorem now follows by adding the first and second term above.