A General Method for Calculating Likelihoods Under the Coalescent Process

K. Lohse,* R. J. Harrison,† and N. H. Barton*,‡,1

*Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom, †East Malling Research, East Malling ME19 6BJ, United Kingdom, and ‡Institute of Science and Technology, A-3400 Klosterneuburg, Austria

ABSTRACT Analysis of genomic data requires an efficient way to calculate likelihoods across very large numbers of loci. We describe a general method for finding the distribution of genealogies: we allow migration between demes, splitting of demes [as in the isolation-with-migration (IM) model], and recombination between linked loci. These processes are described by a set of linear recursions for the generating function of branch lengths. Under the infinite-sites model, the probability of any configuration of mutations can be found by differentiating this generating function. Such calculations are feasible for small numbers of sampled genomes: as an example, we show how the generating function can be derived explicitly for three genes under the two-deme IM model. This derivation is done automatically, using *Mathematica*. Given data from a large number of unlinked and nonrecombining blocks of sequence, these results can be used to find maximum-likelihood estimates of model parameters by tabulating the probabilities of all relevant mutational configurations and then multiplying across loci. The feasibility of the method is demonstrated by applying it to simulated data and to a data set previously analyzed by Wang and Hey (2010) consisting of 26,141 loci sampled from *Drosophila simulans* and *D. melanogaster*. Our results suggest that such likelihood calculations are scalable to genomic data as long as the numbers of sampled individuals and mutations per sequence block are small.

THE coalescent process is highly variable: samples from even a single well-mixed population rapidly coalesce down to a few ancestral lineages, so that their deeper ancestry is determined by just a few random coalescence events (Felsenstein 1992). Thus, small samples taken from a large number of loci give much more information than large samples from a few loci. For example, the distribution of coalescence times, and hence the history of effective population size, has been inferred from single diploid genomes (Li and Durbin 2011). Although it is now feasible to sample very large numbers of markers, or indeed whole genomes, we urgently need methods for analyzing such data. In principle, we can calculate likelihoods from very large data sets, if we have loosely linked blocks of sequence within which recombination is negligible. Provided that only a few genomes are sampled, we can tabulate the probability that any particular configuration of mutations will be seen at each locus and then multiply across large numbers of loci to find the likelihood of our model (Takahata *et al.* 1995).

Wilkinson-Herbots (2008) and Wang and Hey (2010) derive the distribution of coalescence times for a pair of genes sampled from two populations that separated at some time in the past and subsequently exchanged migrants. This "isolation-with-migration" (IM) model is of particular interest in evaluating the role of gene flow during speciation. Hobolth *et al.* (2011) show how this and similar calculations can be done more efficiently using matrix exponentials.

Here, we present an alternative method, based on generating functions, which provides direct information about the pattern of mutational variation and can be automated using symbolic algebra packages such as *Mathematica*. We give the IM model as an example and show how the method extends to linked loci.

The Generating Function of a Genealogy

The ancestry of a sample of genes, Ω, is described by the lengths of the branches that are ancestral to every possible subset. For example, suppose that we have three genes at a locus, labeled $\Omega = \{a, b, c\}$. We label lineages by the set

doi: 10.1534/genetics.111.129569
Manuscript received April 13, 2011; accepted for publication August 18, 2011
Supporting information is available online at http://www.genetics.org/content/suppl/2011/09/07/genetics.111.129569.DC1.

[1] Corresponding author: Institute of Science and Technology, Am Campus 1, A-3400 Klosterneuburg, Austria. E-mail: nick.barton@ist.ac.at

of genes to which they are ancestral. Thus, if lineages ancestral to genes b and c coalesced most recently, then the branches $\{b\}$ and $\{c\}$ have the same length; *i.e.*, $t_{\{b\}} = t_{\{c\}}$, and $t_{\{a\}} = t_{\{b\}} + t_{\{b,c\}}$. With this topology, there are no lineages ancestral to $\{a, b\}$ or $\{a, c\}$ and $t_{\{a,b\}} = t_{\{a,c\}} = 0$. Thus, both the topology and the branch lengths are encoded by the vector of all possible branches $\underline{t}$, which has elements t_S for $S \subseteq \Omega$.

The generating function (GF) for the branch lengths $\underline{t}$ depends on a set of corresponding dummy variables, $\underline{\omega}$ and is defined as the expectation $\psi[\underline{\omega}] = E[e^{-\underline{\omega}.\underline{t}}]$. It is more convenient to use this form—a Laplace transform—rather than the alternative $E[\prod_{S\subseteq\Omega} z_S^{t_S}]$. Generating functions are widely used, primarily because the distribution of the sum of two independent variables is given by the product of the corresponding GF. In particular, Latter (1973) used a GF approach to find the solution for the expected frequency of heterozygotes under the symmetric IM model and Griffiths (1981b) used the GF for the numbers of types to calculate sampling distributions for the infinite-alleles model. Griffiths (1991) applied this to the two-locus problem (see also Jenkins 2008). In the context of the coalescent, the GF has a concrete interpretation: under the infinite-sites model, it is the probability of seeing no mutations, given mutation rate ω_S along branch S.

Information about the branch lengths themselves can be recovered from the GF. The mean lengths, $E[t_S]$, are found by differentiating with respect to ω_S and setting $\underline{\omega}$ to zero; higher moments are found by differentiating more than once. The actual distribution can be found by taking the inverse Laplace transform, which may be done either algebraically (if the GF has a certain form) or by numerical integration.

In practical applications, we wish to know the probability that there are k_S mutations on branch S. Under the infinite-sites model, with mutation rate μ, this is given by taking the expectation of a Poisson distribution with mean μt over the distribution of coalescence times,

$$P[k_S] = E\left[e^{-\mu t_S}\frac{(\mu t_S)^{k_S}}{k_S!}\right] = \frac{(-\mu)^{k_S}}{k_S!}\left(\frac{\partial^{k_S}\psi}{\partial\omega_S^{k_S}}\right)_{\omega_S=\mu}, \tag{1}$$

which is proportional to the k_S'th differential of the GF with respect to ω_S, taken at $\omega_S = \mu$, and setting all other ω's to zero. We see that Equation 1 defines a term in a Taylor series, so that the probability of a particular configuration of mutations is given by the coefficient in the expansion of ψ. In other words, if we set $\omega_S = \mu - x_S$ and expand around the point $x_S = 0$, then the probability of seeing k_S mutations on branch S is the coefficient of $x_S^{k_S}$, multiplied by μ^{k_S}. Similarly, the joint probability of seeing a configuration of $k_{S_1}, k_{S_2}, \ldots$ mutations on branches S_1, S_2, ... is the coefficient of $x_{S_1}^{k_{S_1}} x_{S_2}^{k_{S_2}} \ldots$, multiplied by $\mu^{k_{S_1}+k_{S_2}\cdots}$. In the following, we scale time relative to twice the effective population size, $2N$; *i.e.*, the scaled mutation rate is $2N\mu = \theta/2$.

While we assume an infinite-sites mutation model for simplicity throughout, the GF can also be used to obtain the probabilities of mutational configurations for more complex mutation models. For example, under the Jukes-Cantor (Jukes and Cantor 1969) model mutations to a different state happen at rate $(3/4)\mu$ and the chance of a back mutation is $(1/4)\mu$. The probabilities that two sequences differ or are the same at any particular site are $3(1-e^{-\mu t})/4$ and $(1+3e^{-\mu t})/4$, respectively. Given a pair of sequences of length n the probability of seeing j sites in a different and $n-j$ in the same state is given by taking the expectation of a Binomial distribution over the distribution of coalescence times:

$$P[j] = E\left[\left(\frac{3}{4}\right)^n (1-e^{-\mu t})^j \left(\frac{1}{3}+e^{-\mu t}\right)^{n-j} \binom{n}{j}\right]. \tag{2}$$

This can be written as a sum of the GFs of pairwise coalescence times:

$$P[j] = \left(\frac{3}{4}\right)^n \sum_{k=0}^{j} \sum_{a=k}^{n-j+k} (-1)^k \left(\frac{1}{3}\right)^{n-j-a+k} \binom{n}{j}\binom{j}{k}\binom{n-j}{a-k}\psi[\mu a]. \tag{3}$$

Thus, in principle, we can obtain results under a finite-sites mutation model directly from the GF without the need to take derivatives.

The generating function is a sum of terms, each corresponding to a particular topology. For a given topology, many branches will have zero length by definition, and so the GF will be independent of the corresponding ω_S; some branches will have the same lengths (*e.g.*, $t_{\{b\}} = t_{\{c\}}$) and so the corresponding terms will be a function of the sum of the respective dummy variables (*e.g.*, $\omega_{\{b\}} + \omega_{\{c\}}$). Under the infinite-sites model, this brings a substantial simplification if we see mutations on internal branches, because any terms that do not depend on the corresponding dummy variables can be dropped from the GF: they represent topologies inconsistent with the data. The joint likelihood for a given mutational configuration can then be calculated by multiple differentiation of the remaining terms, which involves a sum over only the possible topologies.

The General Recursion

The recursion for the generating function of genealogical branch lengths can be derived by tracing back from the present to the most recent event, which might be a coalescence, a recombination, a movement between demes, a change in population structure, or whatever. Events i occur at rate λ_i and (tracing back in time) change the configuration of genes from the sampling configuration Ω to Ω_i. Configurations include the number of lineages and—depending on the model—their locations and/or genetic backgrounds. For example, suppose that we start with three lineages $\{a\}$,

$\{b\}$, and $\{c\}$. A coalescence between lineages $\{b\}$ and $\{c\}$ generates a new configuration $\{\{b, c\}, \{a\}\}$, in which there are now two lineages—one ancestral to $\{b, c\}$ and the other to $\{a\}$. We derive a recursion that expresses the GF $\psi[\Omega]$ as a sum over the possible configurations before the previous event. The time back to that event is exponentially distributed with rate $\sum_i \lambda_i$, and so the distribution of the lengths of the terminal branches is just the convolution of this with their previous distribution. Taking Laplace transforms, this corresponds simply to multiplication by the factor $1/(\sum_i \lambda_i + \sum_{|S|=1} \omega_S)$, since a convolution of distributions transforms to a product of the previous GF and the GF of an exponential distribution with rate $\sum_i \lambda_i$. Summing over all possible events we have

$$\psi[\Omega] = \frac{\sum_i \lambda_i \psi[\Omega_i]}{\left(\sum_i \lambda_i + \sum_{|S|=1} \omega_S\right)}. \tag{4}$$

The denominator gives the total rate of events, $\sum_i \lambda_i$ in the interval from the present to the first event, plus the sum of the ω_S that correspond to terminal branches (the “leaves” of the tree). The numerator is the sum over all possible generating functions at the previous event; Ω_i denotes the configuration prior to event i. This recursion yields a set of linear equations for the $\psi[\Omega]$ that is readily solved; the limit is set by the number of possible sample configurations of genes that have to be tracked. To see how this works, we give a series of examples.

A Single Population

In the simplest case of a single well-mixed population, we need to track only coalescence events. Scaling time relative to twice the effective population size, $2N$, the rate of coalescence is given by the number of pairs of lineages in a given sample configuration $\binom{|\Omega|}{2} = |\Omega|(|\Omega|-1)/2,$ where there are $|\Omega|$ lineages. Thus

$$\psi[\Omega] = \frac{1}{\left(\binom{|\Omega|}{2} + \sum_{|S|=1} \omega_S\right)} \sum_{\{x,y\} \subseteq \Omega} \psi\left[\Omega_{\{x,y\}}\right], \tag{5}$$

where the sum is over all the $\binom{|\Omega|}{2}$ possible pairwise coalescences, between genes x and y. $\Omega_{\{x,y\}}$ denotes the sample configuration after coalescence, *i.e.*, Ω with lineages $\{x\}$, $\{y\}$ replaced by the new lineage $\{x, y\}$. Since we define the GF for a single gene as 1, we have for two genes

$$\psi[a, b] = \frac{1}{(1 + \omega_a + \omega_b)}. \tag{6}$$

This is equivalent to the probability of identity in state with $\omega_a + \omega_b = \theta$. Note that for brevity, we have condensed the notation so that $\psi[a, b]$ represents the GF for two lineages ancestral to genes a and b, respectively; and $\psi[ab, c]$ represents two lineages, one ancestral to a and b, and the other to c. For automated recursions (File S1), the full (and unambiguous) notation $\psi[\underline{\omega}, \{\{a\}, \{b\}\}]$, $\psi[\underline{\omega}, \{\{a, b\}, \{c\}\}]$ would be used. For three genes

$$\psi[a, b, c] = \frac{1}{(3 + \omega_a + \omega_b + \omega_c)} \times \left(\frac{1}{(1 + \omega_{ab} + \omega_c)} + \frac{1}{(1 + \omega_{ac} + \omega_b)} + \frac{1}{(1 + \omega_{bc} + \omega_a)}\right). \tag{7}$$

Each of the three terms corresponds to one of the three possible topologies. For example, the last term depends on ω_{bc} and corresponds to coalescence between $\{b\}$ and $\{c\}$, so that the interior branch $t_{bc} > 0$. To find the probability of each topology, we set all the ω_S to zero, and see that each term contributes $\frac{1}{3}$. To find the probability that there are k mutations ancestral to b and c, we differentiate k times with respect to ω_{bc}, set ω_{bc} to equal the scaled mutation rate $\theta/2$ and all other ω_S to zero, and multiply by $(-\theta/2)^k/k!$ (Equation 1). This gives the geometric distribution $(1/3)(2\theta^k/(2 + \theta)^{k+1})$ for $k > 0$, the factor 3 arising because there is a 1/3 probability that b and c coalesce first, allowing mutations of this class to exist. Alternatively, we could set all the ω_S to 0, except for $\omega_{bc} = \theta/2 - x_{bc}$, and then expand around $x_{bc} = 0$; the coefficients of $x_{bc}^{k_{bc}}$ are proportional to the chance of seeing k_{bc} mutations that are ancestral to b and to c. The joint probabilities of other mutational configurations can be found in a similar way.

Migration

Suppose that two populations exchange migrants at a scaled rate $2Nm$. For simplicity, we assume that migration is symmetric and both demes are of the same size (the generalization to more demes, different population sizes, and asymmetric migration is obvious) and that a set Ω_1 of genes is sampled from one deme and Ω_2 from the other. Now, there can be coalescence, which reduces the size of one or the other set, or migration, which transfers a lineage x from one deme to the other creating, for example, new sample configurations $\Omega_{1,+x}$ and $\Omega_{2,-x}$. Thus

$$\begin{aligned}\psi[\Omega_1, \Omega_2] &\\ = &\frac{1}{\left(\binom{|\Omega_1|}{2} + \binom{|\Omega_2|}{2} + 2Nm(|\Omega_1| + |\Omega_2|) + \sum_{S \subseteq \Omega_1, |S|=1} \omega_S + \sum_{S \subseteq \Omega_2, |S|=1} \omega_S\right)} \\ &\times \left(\sum_{\{x,y\} \subseteq \Omega_1} \psi\left[\Omega_{1,\{x,y\}}, \Omega_2\right] + \sum_{\{x,y\} \subseteq \Omega_2} \psi\left[\Omega_1, \Omega_{2,\{x,y\}}\right] \right. \\ &\left. + \ 2Nm \sum_{\{x\} \subseteq \Omega_1} \psi\left[\Omega_{1,-x}, \Omega_{2,+x}\right] + 2Nm \sum_{\{x\} \subseteq \Omega_2} \psi\left[\Omega_{1,+x}, \Omega_{2,-x}\right]\right).\end{aligned} \tag{8}$$

This leads to a set of linear equations that can readily be solved. We need to distinguish only sample configurations where the genes are in different demes, $\psi[a \backslash b]$, or in the same demes, $\psi[a, b \backslash \emptyset]$, say (again, we have condensed the notation; $\emptyset$ represents the empty set, and \ the separation between the two demes). From Equation 8 and using the symmetry of the model,

$$\begin{aligned}\psi[a\backslash b] &= \frac{2Nm}{(4Nm+\omega_a+\omega_b)}(\psi[a,b\backslash\emptyset]+\psi[\emptyset\backslash a,b])\\ &= \frac{4Nm}{(4Nm+\omega_a+\omega_b)}\psi[a,b\backslash\emptyset]\\ \psi[a,b\backslash\emptyset] &= \frac{1}{(1+4Nm+\omega_a+\omega_b)}\\ &\quad\times\ (\psi[ab\backslash\emptyset]+2Nm(\psi[a\backslash b]+\psi[b\backslash a]))\\ &= \frac{1}{(1+4Nm+\omega_a+\omega_b)}(1+4Nm\psi[a\backslash b]).\end{aligned} \tag{9}$$

This has the solution

$$\begin{aligned}\psi[a\backslash b] &= \frac{M}{(M+\omega_a+\omega_b)(1+M+\omega_a+\omega_b)-M^2}\\ \psi[a,b\backslash\emptyset] &= \frac{M+\omega_a+\omega_b}{(M+\omega_a+\omega_b)(1+M+\omega_a+\omega_b)-M^2},\end{aligned} \tag{10}$$

where $M = 4Nm$. Note that the GF is a function only of $\omega_a + \omega_b$, given the constraint $t_a = t_b$. Equation 10 has been previously derived as the probability of identity in state with $\omega_a + \omega_b = \theta$ (Griffiths 1981a, equation 10). Taking the inverse Laplace transform gives the probability of pairwise coalescent times,

$$\begin{aligned}P_{a,b\backslash\emptyset}[t] &= \frac{1}{2}\left(e^{-\lambda_0 t}\left(1-\frac{1}{\lambda_1-\lambda_0}\right)+e^{-\lambda_1 t}\left(1+\frac{1}{\lambda_1-\lambda_0}\right)\right)\\ P_{a\backslash b}[t] &= \frac{M\left(e^{-\lambda_0 t}-e^{-\lambda_1 t}\right)}{\lambda_1-\lambda_0},\end{aligned} \tag{11}$$

where $\lambda_0 = \frac{1}{2}(1+2M-\sqrt{1+4M^2})$ and $\lambda_1 = \frac{1}{2}(1+2M+\sqrt{1+4M^2})$. This result was derived directly by Herbots (1997), using a partial fraction expansion (see Griffiths 1981a; Wilkinson-Herbots 2008, equation 18), but can also be found from the discrete time transition matrix (Wakeley 1996). In fact, λ_0 and λ_1 are the eigenvalues of the symmetric transition matrix Q given by Hobolth *et al.* (2011) with $S_1 = S_2$, $S_{11} = S_{22}$, and $m_1 = m_2$.

Population Splits: The IM Model

Now, suppose that the two populations derive from a single ancestral population T generations ago. Dealing with finite times explicitly leads to complicated expressions (Wang and Hey 2010). However, we can retain the simple form of the GF by taking the Laplace transform with respect to the divergence time, with dummy variable Λ. This has a concrete interpretation, as the expectation over a model in which the divergence time is exponentially distributed with rate Λ, times a normalizing factor Λ. We can either fit this model directly or take the inverse Laplace transform with respect to Λ, to find the GF of the genealogy for a given divergence time T, which we denote P. (More precisely, we take the inverse Laplace transform of $\Lambda^{-1}\psi$, since $\psi = E[e^{-\Lambda T}P] = \int_0^\infty \Lambda e^{-\Lambda T}P dT$.)

The recursion is now

$$\begin{aligned}&\psi[\Omega_1,\Omega_2]\\ &= \frac{1}{\left(\Lambda+\binom{|\Omega_1|}{2}+\binom{|\Omega_2|}{2}+2Nm(|\Omega_1|+|\Omega_2|)+\sum_{S\subseteq\Omega_1,|S|=1}\omega_S+\sum_{S\subseteq\Omega_2,|S|=1}\omega_S\right)}\\ &\quad\times\left(\Lambda\psi[\Omega_1\cup\Omega_2]+\sum_{\{x,y\}\subseteq\Omega_1}\psi\left[\Omega_{1,\{x,y\}},\Omega_2\right]+\sum_{\{x,y\}\subseteq\Omega_2}\psi\left[\Omega_1,\Omega_{2,\{x,y\}}\right]\right.\\ &\quad\left.+\ 2Nm\sum_{\{x\}\subseteq\Omega_1}\psi\left[\Omega_{1,-x},\Omega_{2,+x}\right]+2Nm\sum_{\{x\}\subseteq\Omega_2}\psi\left[\Omega_{1,+x},\Omega_{2,-x}\right]\right).\end{aligned} \tag{12}$$

The additional term $\Lambda\psi[\Omega\cup\Omega]$ represents the replacement of the GF for two separate demes by the GF for a single population, which follows the standard coalescent (see Equation 5). Expression (12) is otherwise identical to Equation 8.

As a simple example, consider two genes,

$$\begin{aligned}\psi[a\backslash b] &= \frac{1}{\Lambda+4Nm+\omega_a+\omega_b}\\ &\quad\times\ (\Lambda\psi[a,b]+2Nm(\psi[\emptyset\backslash a,b]+\psi[a,b\backslash\emptyset]))\\ &= \frac{1}{\Lambda+4Nm+\omega_a+\omega_b}\times\left(\frac{\Lambda}{(1+\omega_a+\omega_b)}+4Nm\psi[a,b\backslash\emptyset]\right)\\ \psi[a,b\backslash\emptyset] &= \frac{1}{\Lambda+1+4Nm+\omega_a+\omega_b}\\ &\quad\times\ (\Lambda\psi[a,b]+\psi[ab\backslash\emptyset]+2Nm(\psi[a\backslash b]+\psi[b\backslash a]))\\ &= \frac{1}{\Lambda+1+4Nm+\omega_a+\omega_b}\left(\frac{\Lambda}{(1+\omega_a+\omega_b)}+1+4Nm\psi[a\backslash b]\right),\end{aligned} \tag{13}$$

which have a solution similar to Equation 10:

$$\begin{aligned}\psi[a\backslash b] &= \frac{1}{1+\omega_a+\omega_b}\\ &\quad\times\frac{\Lambda(1+\Lambda+\omega_a+\omega_b)+M(1+2\Lambda+\omega_a+\omega_b)}{\left(\Lambda+\omega_a+\omega_b+(\Lambda+\omega_a+\omega_b)^2+M(1+2\Lambda+2\omega_a+2\omega_b)\right)}\\ \psi[\emptyset\backslash a,b] &= \frac{1}{1+\omega_a+\omega_b}\\ &\quad\times\frac{\Lambda+\omega_a+\omega_b+(\Lambda+\omega_a+\omega_b)^2+M(1+2\Lambda+\omega_a+\omega_b)}{\left(\Lambda+\omega_a+\omega_b+(\Lambda+\omega_a+\omega_b)^2+M(1+2\Lambda+2\omega_a+2\omega_b)\right)}.\end{aligned} \tag{14}$$

With complete isolation (*i.e.*, $M = 0$), differentiation of these expressions yields the explicit formula for the numbers of pairwise differences in the complete isolation model given by Takahata *et al.* (1995).

For three genes we have

$$\begin{aligned}\psi[a\backslash b,c] &= \frac{1}{\Lambda+1+6Nm+\omega_a+\omega_b+\omega_c}\\ &\quad\times\ (\Lambda\psi[a,b,c]+\psi[a\backslash bc]+2Nm(\psi[\emptyset\backslash a,b,c]\\ &\quad+\ \psi[c\backslash a,b]+\psi[b\backslash a,c]))\\ \psi[\emptyset\backslash a,b,c] &= \frac{1}{\Lambda+3+6Nm+\omega_a+\omega_b+\omega_c}\\ &\quad\times\ (\Lambda\psi[a,b,c]+\psi[\emptyset\backslash a,bc]+\psi[\emptyset\backslash ab,c]+\psi[\emptyset\backslash ac,b]\\ &\quad+\ 2Nm(\psi[a\backslash b,c]+\psi[b\backslash a,c]+\psi[c\backslash a,b])).\end{aligned} \tag{15}$$

Although there are only two types of configuration with three genes, there are three permutations of the first. Thus, in our symmetric model, we have four coupled linear equations, which can be written in matrix form,

$$
\begin{aligned}
\begin{pmatrix} \psi[a\backslash b,c] \\ \psi[b\backslash a,c] \\ \psi[c\backslash a,b] \\ \psi[\varnothing\backslash a,b,c] \end{pmatrix} &= \Lambda\psi[a,b,c]\begin{pmatrix} \gamma_1 \\ \gamma_1 \\ \gamma_1 \\ \gamma_3 \end{pmatrix} \\
&+ \begin{pmatrix} \gamma_1\psi[a\backslash bc] \\ \gamma_1\psi[b\backslash ac] \\ \gamma_1\psi[c\backslash ab] \\ \gamma_3(\psi[\varnothing\backslash a,bc]+\psi[\varnothing\backslash b,ac]+\psi[\varnothing\backslash c,ab]) \end{pmatrix} \\
&+ 2Nm\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \times \begin{pmatrix} \gamma_1\psi[a\backslash b,c] \\ \gamma_1\psi[b\backslash a,c] \\ \gamma_1\psi[c\backslash a,b] \\ \gamma_3\psi[\varnothing\backslash a,b,c] \end{pmatrix},
\end{aligned}
$$

where $\gamma_j = 1/(\Lambda + j + 6Nm + \omega_a + \omega_b + \omega_c)$ and j is the number of pairs that can coalesce given a particular sample configuration.

This has an explicit solution, which we derive in detail in File S1 using a simple symbolic algorithm. If the demes were not equivalent because of asymmetric migration and/or differences in effective population size, then we would need to distinguish configurations such as $\psi[a\backslash b,\, c]$ and $\psi[b,\, c\backslash a]$ and would have eight coupled equations.

With coalescence or population splits alone, the recursions can be solved directly: every event leads back to a simpler configuration, with either fewer lineages or fewer demes. However, with migration, we must solve a set of coupled equations. This is easily done numerically, for specific $\underline{\omega}$, but beyond the simplest cases leads to cumbersome algebraic expressions that cannot readily be differentiated. One way around this problem (which we employ in File S1) is to condition on the topology. Another simplification is to expand the GF in $M = 4Nm$, writing $\psi = \sum_{i=0}^{\infty} M^i \psi_i$. Then, each migration event leads back to a lower-order expression, and we can again find the solution directly. This procedure is equivalent to separating out the GF into a sum of terms, each corresponding to 0, 1, 2, . . . migration events.

In comparison, it is straightforward to obtain results for summaries of the genealogy from the GF. For instance, the distribution of the total number of mutations X can be found by setting all $\underline{\omega}$ to be the same and taking the inverse Laplace transform (see File S1). Similarly, the probability of a particular topology can be found by taking the limit of the ω_S corresponding to internal branches that are incompatible with this topology at infinity with all other ω_S evaluated at zero. For a triplet with sampling configuration $\{a\backslash b,\, c\}$ this gives

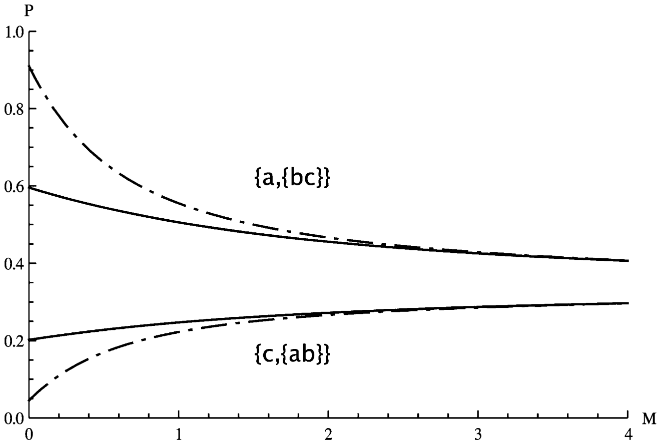


Figure 1 Topological probabilities (Equation 16) for a sample of three genes in the IM model, plotted against the scaled migration rate M for two splitting times, $T = 0.5$ (solid lines) and $T = 2$ (dashed lines). The chance of observing an incongruent genealogy with topology {c, {a, b}} or {b, {a, c}} (bottom) increases with M, as congruent topologies {a, {b, c}} (top) become less likely.

$$
\begin{aligned}
P[\{a,\{b,c\}\}] &= \lim_{\substack{\omega_{ab}\to\infty \\ \omega_{ac}\to\infty}} \psi[a\backslash b,c]\big|_{\omega_S=0} = \frac{2M+3-2e^{-(1+2M)T}}{3(1+2M)} \\
P[\{c,\{a,b\}\}] = P[\{b,\{a,c\}\}] &= \lim_{\substack{\omega_{ab}\to\infty \\ \omega_{bc}\to\infty}} \psi[a\backslash b,c]\big|_{\omega_S=0} = \frac{2M+e^{-(1+2M)T}}{3(1+2M)}.
\end{aligned}
\tag{16}
$$

For the case of three genes in the IM model Equation 16 yields Figure 1. Furthermore, for a given topology, $\{a,\, \{b,\, c\}\}$ say, one can find the distribution of the number of mutations on the internal branch, $P[k_{bc} | \{a,\, \{b,\, c\}\}]$, by differentiating the limit in Equation 16 with respect to ω_{bc} and setting all other ω_S to zero as before. Plotting these distributions (Figure 2) reveals that genealogies congruent with the sampling, *i.e.*, with topology $\{a,\, \{b,\, c\}\}$, tend to have a longer internal branch than those with incongruent topologies $\{b,\, \{a,\, c\}\}$ or $\{c,\, \{a,\, b\}\}$ (Figure 2A *vs.* 2B). This is to be expected, given that coalescence events between lineages sampled from the same population, in this case $\{b,\, c\}$, occur relatively faster, leaving a long time t_{bc} during which mutations can occur on the internal branch. In contrast, coalescence events between lineages sampled from different populations are likely to occur deeper in the past, within the ancestral population. These new results extend previous theory on pairwise coalescence times in the IM model (Wilkinson-Herbots 2008; Wang and Hey 2010) to topologically informative samples. Likewise, it is straightforward to use the GF to extend pairwise results for the IM model beyond the two-deme case. Larger numbers of populations (d) would be incorporated into Equation 9 by an additional term $(d - 1)$; *e.g.*, the rate at which pairs of lineages in different demes are brought together in the same population becomes $M/(d - 1)$.

Recombination Between Linked Loci

The GF method readily extends to multiple linked loci. Each individual is represented as a list, which for each locus gives the set of genes to which it is ancestral; Figure 3 gives an example with three loci. Suppose that we have k individuals, carrying lineages $\underline{\Omega} = \Omega_1, \dots, \Omega_k$,

$$\psi[\underline{\Omega}] = \frac{1}{\left(\binom{k}{2} + 2N \sum_{\alpha\in\mathfrak{R}} r_\alpha + \omega_L\right)} \left(\sum_{1\le i<j\le k} \psi[\underline{\Omega}_{i,j}] + 2N \sum_{\alpha\in\mathfrak{R}} r_\alpha \psi\left[\underline{\Omega}_\alpha\right] \right), \tag{17}$$

where $\omega_L = \sum_{i=1}^{k} \sum_{S\subseteq\Omega_i;|S|=1} \omega_S$, *i.e.*, we need to sum the ω_S leaves over both loci and individuals, $\mathfrak{R}$ is the set of all possible recombination events and r_α is the rate of a recombination of type $\alpha \in \mathfrak{R}$. The first sum on the right in Eq. (4) is over all $\binom{k}{2}$ possible coalescences between the k individuals. At each coalescence, the lists of genes at each locus are merged. For example, a coalescence between $\{a,\ p,\ x\}$ and $\{\emptyset,\ q,\ \emptyset\}$ gives an ancestral lineage $\{a,\ pq,\ x\}$, in which the second locus is now ancestral to genes p and q. The second sum on the right is over all possible recombination events $\alpha \in \mathfrak{R}$, each resulting in a new set of lineages $\underline{\Omega}_\alpha$; these increase the number of lineages to $k + 1$. For example, a recombination in the parent of an individual $\{b,\ q,\ y\}$, between the first locus and the other two, gives two ancestral lineages $\{b,\ \emptyset,\ \emptyset\}$ and $\{\emptyset,\ q,\ y\}$ (Figure 3). Note that this recursion does capture the non-Markovian nature of recombination: the distribution of coalescence times at a locus depends on the genealogies at all the other loci, not just the adjacent locus. The GF gives the joint distribution of genealogies rather than the full ancestral recombination graph (which includes additional information about which loci were carried by the ancestors).

Consider the simplest case, of two genes at two loci; when these are in two individuals, the configuration is denoted $\{a,\ x\}$, $\{b,\ y\}$ and $\omega_{\mathrm{L}} = \omega_{\mathrm{a}} + \omega_{\mathrm{b}} + \omega_{\mathrm{x}} + \omega_{\mathrm{y}}$:

$$\begin{aligned}
\psi[\{a,x\},\{b,y\}] &= \frac{1}{1+4Nr+\omega_L}(\psi[\{ab,xy\}] + 2Nr\psi[\{a,\emptyset\},\{\emptyset,x\},\{b,y\}] \\
&\quad + \psi[\{a,x\},\{b,\emptyset\},\{\emptyset,y\}]) \\
\psi[\{a,\emptyset\},\{\emptyset,x\},\{b,y\}] &= \frac{1}{3+2Nr+\omega_L}(\psi[\{a,x\},\{b,y\}] + \psi[\{a,\emptyset\},\{b,xy\}] \\
&\quad + \psi[\{ab,y\},\{\emptyset,x\}] + 2Nr\psi[\{a,\emptyset\},\{\emptyset,y\},\{b,\emptyset\},\{\emptyset,y\}]) \\
\psi[\{a,\emptyset\},\{\emptyset,x\},\{b,\emptyset\},\{\emptyset,y\}] &= \frac{1}{6+\omega_L}(\psi[\{a,x\},\{b,\emptyset\},\{\emptyset,y\}] + \psi[\{a,\emptyset\},\{\emptyset,x\},\{b,y\}] \\
&\quad + \psi[\{ab,\emptyset\},\{\emptyset,x\},\{\emptyset,y\}] + \psi[\{\emptyset,xy\},\{a,\emptyset\},\{b,\emptyset\}] \\
&\quad + \psi[\{a,y\},\{\emptyset,x\},\{b,\emptyset\}] + \psi[\{b,x\},\{a,\emptyset\},\{\emptyset,y\}]).
\end{aligned} \tag{18}$$

By symmetry, we need only these three recursions, for the cases where the four genes are distributed over two, three, or four individuals. Note that $\psi[\{ab,\ xy\}] = 1$, $\psi[\{a,\ \emptyset\},\ \{b,\ xy\}] = \psi[\{a\},\ \{b\}]$, and so on, connecting these two-locus recursions to the one-locus GF.

This has the solution

$$\begin{aligned}
&\psi[\{a,x\},\{b,y\}] \\
&= \frac{2\left(9+R+6R\phi+R^2\phi\right) + (9+R+2R\phi)\omega_L + \omega_L^2}{18+26R+4R^2+\left(27+19R+2R^2\right)\omega_L + (10+3R)\omega_L^2 + \omega_L^3} \\
&\psi[\{a,\emptyset\},\{\emptyset,x\},\{b,y\}] \\
&= \frac{6+\left(6+13R+2R^2\right)\phi + (1+(7+3R)\phi)\omega_L + \phi\omega_L^2}{18+26R+4R^2+\left(27+19R+2R^2\right)\omega_L + (10+3R)\omega_L^2 + \omega_L^3} \\
&\psi[\{a,\emptyset\},\{\emptyset,x\},\{b,\emptyset\},\{\emptyset,y\}] \\
&= \frac{4+\left(7+13R+2R^2\right)\phi + (8+3R)\phi\omega_L + \phi\omega_L^2}{18+26R+4R^2+\left(27+19R+2R^2\right)\omega_L + (10+3R)\omega_L^2 + \omega_L^3},
\end{aligned} \tag{19}$$

where $\phi = 1/(1+\omega_a+\omega_b) + 1/(1+\omega_x+\omega_y)$, and $R = 2Nr$. These formulas correspond to those previously obtained by Simonsen and Churchill (1997), using a Markov chain method. For example, the covariance of coalescence times between two loci is

$$\mathrm{Cov}\left[T_{ab}, T_{xy}\right] = E\left[T_{xy}T_{ab}\right] - E\left[T_{ab}\right]E\left[T_{xy}\right], \tag{20}$$

which can be found straightforwardly from the GF by taking derivatives with respect to ω_a and ω_x and evaluating at $\omega = 0$, noting that $E[T_{ab}] = E[T_{xy}] = 1$:

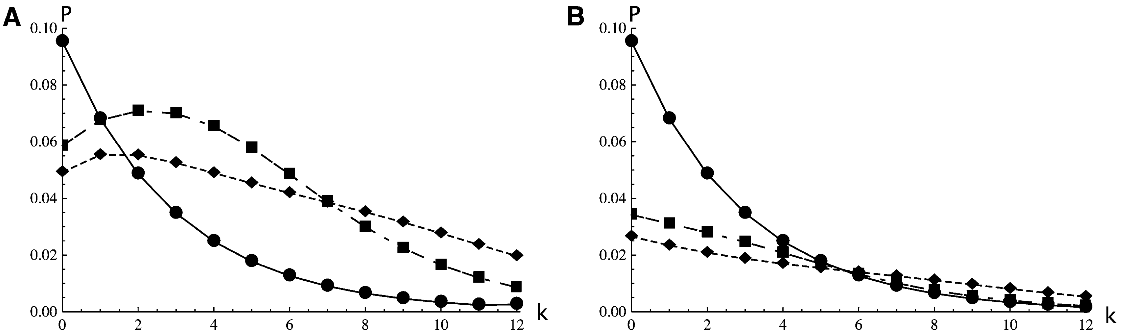


Figure 2 The distribution of the number of mutations (k) on the internal branches for a sample of three genes {a, {b, c}} in the IM model with symmetric migration $\theta = 5$, $M = 0.8$ plotted for three different splitting times $T = 0$ (circles, solid line), $T = 2$ (squares, long-dashed line), and $T = 4$ (diamonds, short-dashed line). Congruent genealogies with topology {a, {b, c}} (A) tend have longer internal branches than those with incongruent topologies {c, {a, b}} or {b, {a, c}} (B). Note that for $T = 0$ the distributions for the two topologies are identical as expected in a panmictic population.

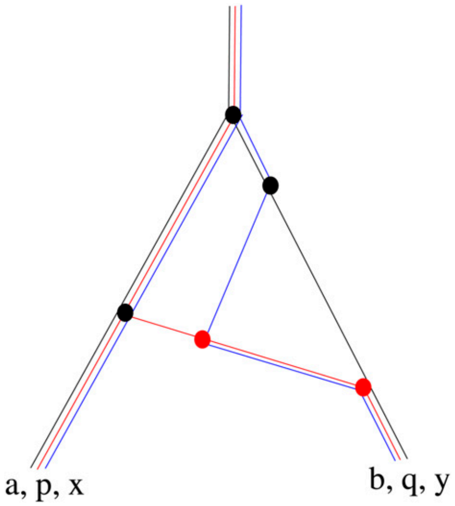


Figure 3 An example of coalescence and recombination between three loci. At the present generation (bottom), there are two individuals: one carries genes *a*, *p*, *x* and the other carries *b*, *q*, *y*. Lineages ancestral to the three loci are colored black, red, and blue, respectively. This is denoted as {{*a*, *p*, *x*}, {*b*, *q*, *y*}}. Tracing back, the most recent event is a recombination (red dot) giving three individuals {{*a*, *p*, *x*}, {*b*, ø, ø}, {ø, *q*, *y*}}, where ø is the empty set. There is then another recombination event, preceded by three coalescence events (black dots); these produce the configurations {{*a*, *p*, *x*}, {*b*, ø, ø}, {ø, *q*, ø}, {ø, ø, *y*}}; {{*a*, *pq*, *x*}, {*b*, ø, ø}, {ø, ø, *y*}}; {{*a*, *pq*, *x*}, {*b*, ø, *y*}}; and {{*ab*, *pq*, *xy*}}. Recombination and coalescence events prior to this single common ancestor do not affect the observed genealogy.

$$\mathrm{Cov}\left[T_{ab}, T_{xy}\right] = \left(\left.\frac{d^2\psi[\{a,x\},\{b,y\}]}{d\omega_a d\omega_x}\right|_{\omega=0}\right) - 1 = \frac{9+R}{9+13R+2R^2}. \tag{21}$$

This agrees with Simonsen and Churchill (1997, equation 52).

Including recombination leads to sets of coupled linear equations, whose solution involves an unwelcome matrix inversion. As with migration, this problem can be avoided by expanding in powers of *R*, which is equivalent to summing over histories that involve 0, 1, ... recombination events. Moreover, these recombination events are uniformly distributed across the genetic map, and so we have a description of the ancestry of the whole genome and not just of two linked loci. The recursions give us the probability that there are no recombination events, that there is one event producing two blocks with different genealogies, that there are three events producing three blocks of genome, and so on. This may allow likelihoods to be calculated for short sequence blocks, provided that *R* is small.

Slatkin and Pollack (2006) calculate the probabilities of alternative topologies for genes at two loci in three completely isolated species; their recursion is essentially the same as ours, but tracks just the distribution of topologies rather than the full distribution of coalescence times. Since no coalescence can occur until two of the genes are brought together in the same ancestral population prior to the most recent speciation event, this reduces to the case of three linked pairs of genes in two completely isolated species. This case can be solved by the above method, by including a rate of population splits, Λ, which corresponds to the time, *T*, between the two speciation events.

Drosophila melanogaster–D. simulans Divergence

To illustrate the feasibility of the GF method for inference in practice, we applied it to both real and simulated data. We first reanalyzed the genomic data set of *Drosophila melanogaster–D. simulans* compiled and analyzed by Wang and Hey (2010), using a likelihood method for pairwise samples. The data (kindly provided by Y. Wang) consist of alignments of 30,247 blocks of intergenic sequence of 500 bp each sampled from two inbred lines of *D. simulans* and one inbred line each of *D. melanogaster* and *D. yakuba* (the latter used as an outgroup to account for mutational heterogeneity and, in the triplet analysis, to polarize mutations). Following Wang and Hey (2010), low-quality sequences, indels, and positions next to indels were removed. Rather than using the divergence to the outgroup to scale the mutation rate at each locus (Yang 2002; Wang and Hey 2010), each locus was trimmed after a fixed number of mutational differences between *D. yakuba* and *D. melanogaster.* We chose a cutoff of 16 divergent sites, which corresponds roughly to a third of the observed mean divergence across all loci in the full data set. A total of 2,090 loci that were below this cutoff were excluded from the analysis. Since our method assumes infinite-sites mutations, sites with more than two segregating states (12.9% of all polymorphic sites) were excluded. We also filtered out shared derived mutations that were topologically incongruent with the majority class of shared derived mutations in each block (2.5% of all polymorphic sites). A total of 2,016 loci, which contained equal numbers of topologically conflicting shared derived mutations, were excluded. The final, trimmed data set consisted of 26,141 loci. To convert scaled parameter estimates into absolute values ($N_{\mathrm{e}} = \theta/4\mu$, $t = g2N_{\mathrm{e}}T$), we followed Wang and Hey (2010) and assumed that *D. yakuba* and *D. melanogaster* split 10 MYA and with a generation time per year of $g = 0.1$, which gives a mutation rate per block of 8×10^{-8}.

Given that Wang and Hey (2010) detected a signal of gene flow from *D. simulans* to *D. melanogaster* but not in the reverse direction, we fitted an IM model with asymmetric migration. The GF for this case can be obtained using Equation 4 and, given that each genealogy can be affected by only one migration event at most, is considerably simpler than the analogous expression with symmetric migration given by solving Equation 15 (details are provided in File S1). To investigate the effect (in terms of bias and power) of including a third sample and thus topology information on parameter estimation, we performed analogous likelihood analyses on pairwise (one sample from each of *D. melanogaster* and *D. simulans*) and triplet data. To assess the effect of removing positions that violate the infinite-sites mutation model, we also ran a pairwise analysis on the full, untrimmed data set. Mutational heterogeneity in this

Table 1 Population parameters estimated for *D. melanogaster*–*D. simulans* using 26,141 loci (data from (Wang and Hey 2010)

Data set	θ (N_e)	$M = 4Nm$	T (t)	logL
Pair, full data	1.85 (5.52×10^6)	0.051	2.70 (2.98×10^6)	−93,466
Pair, trimmed[a]	**1.51 (**4.72×10^6**)**	**0.093**	**3.34 (**3.15×10^6**)**	−65,717
Triplet, trimmed[a]	1.40 (4.37×10^6)	0.174	3.34 (2.92×10^6)	−149,556
Pair, simulated	1.53 (4.79×10^6)	0.098	3.24 (3.10×10^6)	−65,619
Triplet, simulated	1.51 (4.72×10^6)	0.092	3.29 (3.11×10^6)	−151,483

Absolute values are in parentheses. MLEs for M and t in the pairwise analysis agree well with the results of Wang and Hey (2010) who estimated $t = 3.04$ and $M = 0.059$ (after correction for differences in scaling M). The filtering necessary to satisfy the infinite-sites model leads to a decrease in the estimate of N_e and an increase in M. The last two rows show parameters estimated from data simulated using the MLE from the pairwise analysis (boldface type).

[a] Trimmed refers to shortening each locus to a fixed outgroup divergence and removing back mutations and topologically incongruent mutations.

analysis was incorporated by binning loci according to their outgroup divergence and specifying mutation rate scalars for each bin (we used 10 bins).

To speed up calculations in the triplet analysis the GF was conditioned on the topology (by taking limits as shown in Equation 16). Probabilities of all observed mutational configurations were tabulated separately for each topology class (congruent, incongruent, and topologically uninformative loci) (see File S1). Using the FindMaximum function in *Mathematica*, the joint likelihood of M, T, and θ can be maximized very efficiently (a few seconds or minutes for pairs or triplets, respectively). A *Mathematica* notebook for this calculation is provided in File S1; scripts for preprocessing input data are available from the authors on request.

Despite the fact that we are assuming an infinite-sites mutation model [Wang and Hey (2010) used a Jukes–Cantor (Jukes and Cantor 1969) model], the results from the pairwise analysis on the full data (Table 1) agree well with those obtained by Wang and Hey (2010). As expected, our maximum-likelihood estimate (MLE) of N_e (5.5×10^6) falls in between the N_e estimates obtained by Wang and Hey (2010, Table 7) for the ancestral population (3.1×10^6) and *D. simulans* (5.9×10^6) (note that Wang and Hey 2010 fit a slightly more complex history with separate N_e parameters for each species). Likewise, estimates of M and T agree well with the results of Wang and Hey (2010). The trimming of back mutations and topologically incongruent mutations led to a slight decrease in N_e and increased estimates of M in the pairwise analysis. This effect was more pronounced in the triplet analysis; in particular, the MLE for M was threefold higher than the estimate of Wang and Hey (2010) (Table 1). Furthermore (and perhaps unexpectedly) we found no increase in power in the triplet analysis (Figure 4). To investigate this further, we repeated these analyses on simulated data generated using ms (Hudson 2002) under the IM history estimated for the two *Drosophila* species, *i.e.*, using the MLE obtained from the pairwise analysis on the trimmed data (Table 1). In contrast to the *Drosophila* analyses, we found no bias in parameter estimates and higher power to estimate M and T in triplet compared to pairwise analyses of these simulated data (Figure 4). This suggests that the differences between pairwise and triplet analyses seen in the *Drosophila* example result from violations of the infinite-sites mutation model rather than from an inherent bias of our method. An obvious interpretation is that the use of shared derived mutations to infer the topology at each locus in the triplet analysis makes our method sensitive to misinference of ancestral states resulting from backmutations on the outgroup branch. In other words, mispolarized mutations artificially inflate the proportion of loci with incongruent topologies and hence the estimate of M. As a simple check, we can ask what the expected frequencies of congruent, incongruent, and topologically uninformative loci are (these can be derived from the GF analogous to Equation 16; see File S1). Given the MLE for trimmed pairwise and triplet analysis (Table 1), we expect 2.1% incongruent and 15.7% topologically uninformative loci on the basis of the pairwise results and 2.6% incongruent and 19.3% uninformative loci on the basis of the triplet results. However, the observed frequencies in the data set are 6.2% and 18.8% for topologically incongruent and uninformative loci, respectively. This confirms that there is an apparent (and likely artificial) excess of incongruent topologies in the data that explains the bias seen the triplet MLEs. While this illustrates the problems of assuming infinite-sites mutations when dealing with old divergence events, it is actually surprising how little effect ignoring back mutations had in this case, considering the large distance between in- and outgroup.

We also analyzed triplet data simulated under the reverse sampling scheme (two individuals from the species/population receiving migrants). The GF for this is slightly more complicated and is derived in File S1. The power to estimate M in this case increases substantially when analyzing triplets (Figure 4). This is expected given that most migration events will result in incongruent genealogies with relatively long internal branches.

Discussion

The GF framework provides a general method to derive likelihoods under a variety of models that include migration, changes in population structure, and recombination and applies to arbitrary sample sizes. Here our aim is to set out the method and show that it can be implemented for indefinitely large numbers of loci. So, we have focused on small samples for simplicity. Assuming that populations are exchangeable in size and rate of migration reduces both the number of parameters to be estimated and the number of configurations to track. In the case of the symmetric IM model, we do not need to distinguish the two demes, which

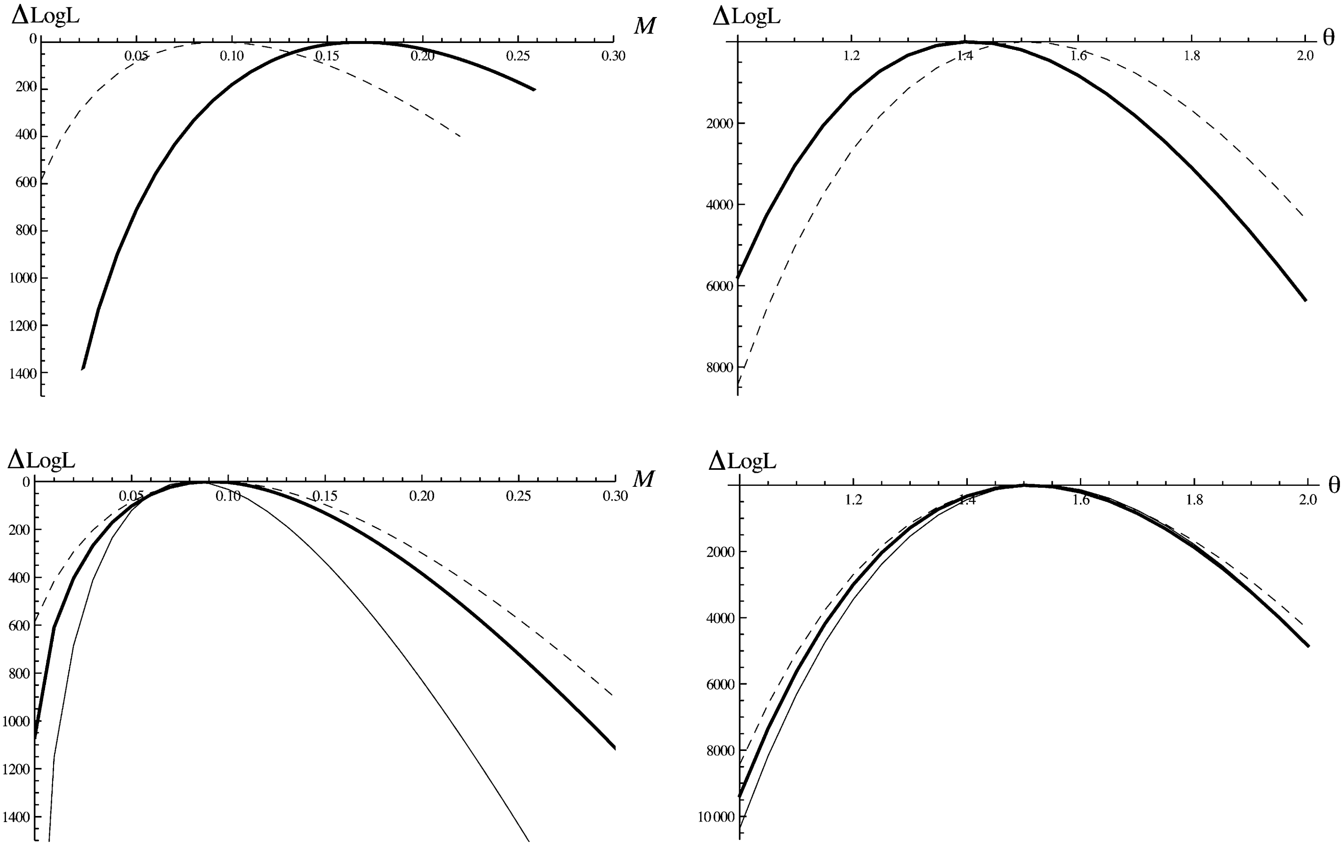


Figure 4 Profile log-likelihood curves for *M* (left plots) and θ (right plots) for pairwise (dashed lines) and triplet analyses (thick solid lines) calculated from 26,141 loci for *D. melanogaster* and *D. simulans* (Wang and Hey 2010) (top row) and simulated data under an IM model with migration from *D. simulans* to *D. melanogaster* (bottom row). Analysis of the *Drosophila* data suggests an apparent bias of the triplet MLE of *M* and no improvement in power. Comparison with data simulated under the same history (using the MLE obtained in the pairwise analysis, see Table 1) shows no bias and tighter log-likelihood for the triplet analyses as expected. The improvement in power when adding a third individual is greater if this is sampled from the species receiving migrants [*i.e.*, the reverse sampling as in the *Drosophila* example (thin solid lines)].

halves the number of sample configurations. At the opposite extreme, under a highly asymmetric model with unidirectional migration (as in the *Drosophila* example above), each lineage in the receiving population can be affected by only a single migration event at most, which also greatly simplifies the problem. More generally, although it is possible to calculate the GF for fairly complex problems (up to six genes in the IM model, say), it is harder to extract useful information from it. Thus, while we can readily find the properties of chosen summary statistics (for example, the number of segregating sites), tabulating the probability of all observed mutational configurations is limited by their sheer number, rather than by the difficulty of finding the GF itself. These computational issues are explored in File S1, using automated recursions for the IM model with three genes.

Our GF approach is more flexible than those of Wang and Hey (2010) and Hobolth *et al.* (2011) in two ways. First, the recursions for a given data set can be simplified by dropping terms that are incompatible with the observed mutational pattern. This strategy is closely related to importance sampling schemes (*e.g.*, Griffiths and Tavaré 1994). Thus, instead of summing over all possible topologies, the calculation is reduced to histories that are possible, given the data. For a sample with a fully resolved topology, the total number of terms is given by the number of configurations due to migration, so that for $n = 4$ and 6 there are only 28 and 124 configurations, respectively. Thus, solutions at least for symmetric cases are feasible. Second, other processes, such as recombination or changes in population size, can easily be incorporated into the GF framework. Since, under the IM model, genealogies involving migration events tend to be shorter and thus more likely to be shared between linked loci, incorporating recombination should improve inference.

Given that species may diverge gradually in space and/or ecology, it makes sense to model population separation as an explicit process, rather than an instantaneous event, followed by constant gene flow. We must distinguish here between our GF method, which calculates an average over exponentially distributed split times, and more general models that allow varying rates of gene flow. We follow the IM model in assuming that populations split abruptly and that subsequently, genes flow at a constant rate. Our initial assumption of an exponential distribution of separation times (with rate Λ) can be viewed either as a technical ruse to allow us to recover the distribution at a specific time, T, by taking an inverse Laplace transform or, in Bayesian terms, as expressing our prior beliefs about T. In reality, gene flow is likely to decrease gradually as populations diverge, and we can imagine a variety of models for the way rates of gene flow vary through time. However, even with large data sets there may be little power to detect changes in the rate of gene flow (Becquet and Przeworski 2009); the question of whether rates of gene flow vary across loci as a result of selection is yet more challenging, but crucial to identifying genes responsible for reproductive isolation (*e.g.*, Machado 2002).

Yang (2010) recently introduced a model that is related to both approaches just described. This assumes that populations separate suddenly, with no subsequent gene flow, but that the split time varies across loci, following a beta distribution—which can be regarded as an approximation to a biologically feasible model in which migration causes variation in coalescence time across loci. This is related to, but different from, our assumption of an exponential rate, Λ, of separation times. If, following Yang (2010), we assumed exponentially distributed split times across loci, we would fix Λ to find the probability of mutational configurations. On the other hand, if we assumed a definite separation time T, we would take the inverse Laplace transform at T and calculate the probabilities from that. If we then averaged the multilocus likelihood over a prior distribution of T, we would get a quite different result from that yielded by Yang's (2010) procedure.

As our application to the *Drosophila* data demonstrates, the GF method outlined here provides an efficient way to calculate and maximize the joint likelihood of divergence parameters from very many nonrecombining blocks of sequence for topologically informative samples. Not only do triplet samples (as opposed to pairs) give better information about branch lengths but also, more importantly, the joint distribution of topologies and branch lengths provides qualitatively new information about historical parameters. As our simulation example demonstrates, dependent on the sampling scheme, this substantially increases power. Our analytic solutions have three key advantages over previous methods. First, the probabilities of mutational configurations need to be tabulated only once, so in contrast to simulation-based methods computation time does not increase with the number of loci and an indefinite number of loci can be analyzed. Second, derivatives can be used to maximize the joint log-likelihood, which greatly speeds up calculations. Thus our computation takes a fraction of the time of, for example, an IMa analysis (Hey and Nielsen 2004) on a handful of loci and is also more efficient than the numerical method of Wang and Hey (2010) (Y. Wang, personal communication). Finally, the GF method allows us to separate topology and branch length information, which provides a way to incorporate additional sources of information. For example, topology information contained in the patterns of shared derived indels could be included without the need to model indel evolution explicitly.

In practice, however, our method is currently limited to the infinite-sites mutation model and thus can deal with only relatively recent divergence events for which close outgroups are available. However, it is encouraging how small the bias resulting from assuming infinite-sites mutations is in the *Drosophila* example, despite the considerable divergence of the outgroup. Fortunately, researchers are commonly interested in fitting IM histories to sister taxa or populations that have diverged much more recently than the *Drosophila* species analyzed here (and for which more closely related outgroups are available). The use of multiple outgroups to correct for misinferred ancestral states should also help to overcome this problem. Another limitation is that the GF can be used to find exact solutions only if the number of mutations per genealogical branch is relatively small (*e.g.*, the most diverse locus in the trimmed *Drosophila* data set contained 26 mutations). For much larger numbers of mutations per block, numerical calculations, which involve finding the coefficients in a series expansion, become unfeasible. Although it may be possible to use a Gaussian approximation in this case, the assumption of no recombination within blocks restricts our and related methods (Hey and Nielsen 2004; Wang and Hey 2010) to short blocks of sequence anyway, so this may not be relevant in practice.

Implementing efficient inference schemes for biologically realistic histories clearly requires further work. For instance, it would be worthwhile to extend our inference scheme to the general IM model (*i.e.*, allowing for asymmetric migration in both directions and different population sizes) and more realistic mutation models and incorporate recombination explicitly. In contrast, the catastrophic increase of possible sample and mutational configurations with the number of individuals frustrates full results for large numbers of individuals. Nevertheless, full results for small but topologically informative samples under a range of models of structure and history should be of considerable interest for at least three reasons: first, although thorough investigations of the trade-offs of various sampling schemes are lacking, it is clear that in general replication across loci is far more profitable than analyzing a few loci sampled from a large number of individuals (Felsenstein 1992; Li and Durbin 2011). Second, minimal sampling in terms of individuals reflects the practical limitations of current sequencing technologies. While massively paralleled sequencing has made it affordable to sequence small numbers of genomes in any organism, obtaining multilocus sequence data for many individuals remains challenging in nonmodel organisms. Finally, under a wide range of models of population structure, large samples quickly coalesce down to a few lineages that dominate their genealogical history, allowing a separation of timescales to be applied (Wakeley 2009). Thus, we envisage that new analytic solutions of simple cases, such as those derived here for the total number of mutations and topological probabilities of triplets under the IM model, will provide a guide to the development of approximate methods (involving importance sampling and summary statistics) with wide applicability.

Acknowledgments

We thank Yong Wang and Jody Hey for sharing the *Drosophila* data set. We also thank two anonymous reviewers and Jerome Kelleher for thoughtful comments on earlier versions of this manuscript. This work was supported by a grant from the European Research Council (250152) (to N.B.) and a grant from the United Kingdom Natural Environment Research Council (NE/I020288/1) (to K.L.).

Literature Cited

Becquet, C., and M. Przeworski, 2009 Learning about modes of speciation from computational approaches. Evolution 63(10): 2547–2562.

Felsenstein, J., 1992 Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. 59: 139–147.

Griffiths, R. C., 1981a The number of heterozygous loci between two randomly chosen completely linked sequences of loci in two subdivided population models. J. Math. Biol. 12: 251–261.

Griffiths, R. C., 1981b Transient distribution of the number of segrating sites in a neutral infinite-sites model with no recombination. J. Appl. Probab. 18: 42–51.

Griffiths, R. C., 1991 The two-locus ancestral graph, pp. 100–117 in, editors, *Selected Proceedings of the Symposium of Applied Probability*, edited by I. V. Basawa and R. I. Taylor. Institute of Mathematical Statistics, Haywards, CA.

Griffiths, R. C., and S. Tavaré, 1994 Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. Lond. B Biol. Sci. 344(1310): 403–410.

Herbots, H., 1997 The structured coalescent, pp. 231–255 in *Progress in Population Genetics and Human Evolution* (IMA Volumes in Mathematics and Its Applications, No. 87), edited by P. Donelly and S. Tavare. Springer-Verlag, Berlin/Heidelberg, Germany/New York.

Hey, J., and R. Nielsen, 2004 Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167: 747–760.

Hobolth, A., L. N. Andersen, and T. Mailund, 2011 On computing the coalescent time density in an isolation-with-migration model with few samples. Genetics 187: 1241–1243.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.

Jenkins, P. A., 2008 Importance sampling on the coalescent with recombination. Ph.D. Thesis, Oxford University, Oxford.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–123 in *Mammalian Protein Metabolism*, edited by H. N. Munro. Academic Press, New York.

Latter, B. D. H., 1973 The island model of population differentiation: a general solution. Genetics 73: 147–157.

Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature 475(7357): 493–496.

Machado, C. A., 2002 Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. Mol. Biol. Evol. 19: 472–488.

Simonsen, K. L., and G. A. Churchill, 1997 A Markov chain model of coalescence with recombination. Theor. Popul. Biol. 52: 43–59.

Slatkin, M., and J. L. Pollack, 2006 The concordance of gene trees and species trees at two linked loci. Genetics 172: 1979–1984.

Takahata, N., Y. Satta, and J. Klein, 1995 Divergence time and population size in the lineage leading to modern humans. Theor. Popul. Biol. 48: 198–221.

Wakeley, J., 1996 Pairwise differences under a general model of subdivision. J. Genet. 75(1): 81–89.

Wakeley, J., 2009 *Coalescent Theory.* Roberts & Co., Greenwood Village, CO.

Wang, Y., and J. Hey, 2010 Estimating divergence parameters with small samples from a large number of loci. Genetics 184: 363–373.

Wilkinson-Herbots, H. M., 2008 The distribution of the coalescence time and the number of pairwise nucleotide differences in the “isolation with migration” model. Theor. Popul. Biol. 73(2): 277–288.

Yang, Z., 2002 Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics 162: 1811–1823.

Yang, Z., 2010 A likelihood ratio test of speciation with gene flow using genomic data. Genome Biol. Evol. 2: 200–211.

Communicating editor: Y. S. Song

GENETICS

Supporting Information
http://www.genetics.org/content/suppl/2011/09/07/genetics.111.129569.DC1

A General Method for Calculating Likelihoods Under the Coalescent Process

K. Lohse, R. J. Harrison, and N. H. Barton

DOI: 10.1534/genetics.111.129569

Supplementary Information

It is easiest to view this document in *Mathematica* or MathPlayer (available as a free download at http://www.wolfram.com/producs/-player/).

1. Automation for the IM model: Three genes in two demes

1.1 Set up

Notation

Lineages are labelled by the set of genes to which they are ancestral. Thus, lineages at the tips are ancestral to a single gene, and are labelled $\{a\}$, $\{b\}$, …. A deme containing lineages $\{b\}$ and $\{c\}$ is denoted $\{\{b\}, \{c\}\}$, and two demes - one containing lineage $\{a\}$ and the other containing $\{b\}$ and $\{c\}$ - is denoted $\{\{\{a\}\}, \{\{b\}, \{c\}\}\}$. If populations can split, we also need to define the ancestry of the demes in a similar way. $\{\{x\}, \{y\}\}$ denotes two demes, ancestral to the present-day demes x and y. The single ancestral deme that existed before the split is denoted $\{\{x, y\}\}$. Note that a single lineage must be ancestral to every gene, and a single deme must be ancestral to every present-day deme. Thus, the content of the lists that define the genealogy and the population phylogeny stays the same - only the nesting changes.

The generating function has the form GF[ω, $\{\{\{a\}\}, \{\{b\}, \{c\}\}\}$, M, $\{\{x\}, \{y\}\}$, Λ]. $\omega[\{a\}]$ corresponds to branch $\{a\}$, which is ancestral to a; $\Lambda[\{x, y\}]$ is the split rate of population $\{x, y\}$. $M = 4\,N\,m$ is the scaled migration rate

In the text, this is denoted more compactly as $\psi[a, b \backslash c]$. tidyNotation[ψ] gives something like this notation, to make the output more readable.

Solving the recursions

This procedure is simple, but not very efficient given that it does not exploit all the symmetries, which can drastically reduce the number of equations needed. However, this part is extremely fast relative to later steps.

makeAllEqns automates the recursions for the IM model. Here we assume a sampling configuartion {a/b,c}.

```
eqs = makeAllEqns[GF[ω, {{{a}}, {{b}, {c}}}, M, {{x}, {y}}, Λ]]; vars = GetVars[eqs]
```

```
{GF[ω, {{{a}, {b, c}}}, M, {{x, y}}, Λ], GF[ω, {{{b}, {a, c}}}, M, {{x, y}}, Λ],
 GF[ω, {{{c}, {a, b}}}, M, {{x, y}}, Λ], GF[ω, {{{a}, {b}, {c}}}, M, {{x, y}}, Λ],
 GF[ω, {{}, {{a}, {b, c}}}, M, {{x}, {y}}, Λ], GF[ω, {{}, {{b}, {a, c}}}, M, {{x}, {y}}, Λ],
 GF[ω, {{}, {{c}, {a, b}}}, M, {{x}, {y}}, Λ], GF[ω, {{}, {{a}, {b}, {c}}}, M, {{x}, {y}}, Λ],
 GF[ω, {{{a}}, {{b, c}}}, M, {{x}, {y}}, Λ], GF[ω, {{{a}}, {{b}, {c}}}, M, {{x}, {y}}, Λ],
 GF[ω, {{{b}}, {{a, c}}}, M, {{x}, {y}}, Λ], GF[ω, {{{b}}, {{a}, {c}}}, M, {{x}, {y}}, Λ],
 GF[ω, {{{c}}, {{a, b}}}, M, {{x}, {y}}, Λ], GF[ω, {{{c}}, {{a}, {b}}}, M, {{x}, {y}}, Λ],
 GF[ω, {{{a, b}}, {{c}}}, M, {{x}, {y}}, Λ], GF[ω, {{{a, c}}, {{b}}}, M, {{x}, {y}}, Λ],
 GF[ω, {{{b, c}}, {{a}}}, M, {{x}, {y}}, Λ], GF[ω, {{{a}, {b}}, {{c}}}, M, {{x}, {y}}, Λ],
 GF[ω, {{{a}, {c}}, {{b}}}, M, {{x}, {y}}, Λ], GF[ω, {{{a}, {b, c}}, {}}, M, {{x}, {y}}, Λ],
 GF[ω, {{{b}, {c}}, {{a}}}, M, {{x}, {y}}, Λ], GF[ω, {{{b}, {a, c}}, {}}, M, {{x}, {y}}, Λ],
 GF[ω, {{{c}, {a, b}}, {}}, M, {{x}, {y}}, Λ], GF[ω, {{{a}, {b}, {c}}, {}}, M, {{x}, {y}}, Λ]}
```

Next, we choose those equations that involve 1 deme, and solve them. First/@eqs1 lists the GF[] that we need to solve for:

```
eqs1 = selectEqns[eqs, {1, All}];
soln1 = Solve[eqs1, First /@ eqs1][[1]]
```

$$
\begin{aligned}
\Big\{&\text{GF}[\omega,\{\{\{a\},\{b,c\}\}\},M,\{\{x,y\}\},\Lambda]\to-\frac{1}{-1-\omega[\{a\}]-\omega[\{b,c\}]},\\
&\text{GF}[\omega,\{\{\{b\},\{a,c\}\}\},M,\{\{x,y\}\},\Lambda]\to-\frac{1}{-1-\omega[\{b\}]-\omega[\{a,c\}]},\\
&\text{GF}[\omega,\{\{\{c\},\{a,b\}\}\},M,\{\{x,y\}\},\Lambda]\to-\frac{1}{-1-\omega[\{c\}]-\omega[\{a,b\}]},\\
&\text{GF}[\omega,\{\{\{a\},\{b\},\{c\}\}\},M,\{\{x,y\}\},\Lambda]\to\\
&\quad 1/((3+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}])\,(1+\omega[\{c\}]+\omega[\{a,b\}]))-\\
&\quad 1/((3+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}])\,(-1-\omega[\{b\}]-\omega[\{a,c\}]))-\\
&\quad 1/((3+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}])\,(-1-\omega[\{a\}]-\omega[\{b,c\}]))\Big\}
\end{aligned}
$$

We then choose those that involve 2 genes in 2 demes:

```
eqs2 = selectEqns[eqs, {2, 2}];
soln2 = Solve[eqs2, First /@ eqs2]⟦1⟧;
```

This needs to be simplified, by using the solutions for all the 1-deme cases (stored in soln1). This is the solution for all configurations with two genes in two demes. Note that this is inefficient: there are 12 configurations in general, but only three kinds for the symmetric model (where both demes have equal popultaion size and migration is symmetric) - the genes can be in the same deme or different demes.

These are the solutions for two genes with two demes, given in the "tidy notation". $\psi_{\{\},\{\{a\},\{b,c\}\}}$ denotes an empty deme, and a deme containing two lineages - one ancestral to $\{a\}$, the other to $\{b, c\}$.

```
soln2Simp = soln2 /. soln1 // Simplify;
soln2Simp /. tidyNotation[ψ] /. {ω{x_, y_} :> ωL - ωComplement [{a,b,c},{x,y}], Λ{x},{y} → Λ} // Simplify
```

$$
\begin{aligned}
\Big\{&\psi_{\{\},\{\{a\},\{b,c\}\}}\to\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(1+M+2\,\Lambda)\,\omega_L+\omega_L^2\right)\Big/\\
&\quad\left((1+\omega_L)\,\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(1+2\,M+2\,\Lambda)\,\omega_L+\omega_L^2\right)\right),\ \psi_{\{\},\{\{b\},\{a,c\}\}}\to\\
&\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(1+M+2\,\Lambda)\,\omega_L+\omega_L^2\right)\Big/\left((1+\omega_L)\,\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(1+2\,M+2\,\Lambda)\,\omega_L+\omega_L^2\right)\right),\\
&\psi_{\{\},\{\{c\},\{a,b\}\}}\to\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(1+M+2\,\Lambda)\,\omega_L+\omega_L^2\right)\Big/\\
&\quad\left((1+\omega_L)\,\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(1+2\,M+2\,\Lambda)\,\omega_L+\omega_L^2\right)\right),\ \psi_{\{\{a\}\},\{\{b,c\}\}}\to\\
&\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(M+\Lambda)\,\omega_L\right)\Big/\left((1+\omega_L)\,\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(1+2\,M+2\,\Lambda)\,\omega_L+\omega_L^2\right)\right),\\
&\psi_{\{\{b\}\},\{\{a,c\}\}}\to\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(M+\Lambda)\,\omega_L\right)\Big/\\
&\quad\left((1+\omega_L)\,\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(1+2\,M+2\,\Lambda)\,\omega_L+\omega_L^2\right)\right),\ \psi_{\{\{c\}\},\{\{a,b\}\}}\to\\
&\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(M+\Lambda)\,\omega_L\right)\Big/\left((1+\omega_L)\,\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(1+2\,M+2\,\Lambda)\,\omega_L+\omega_L^2\right)\right),\\
&\psi_{\{\{a,b\}\},\{\{c\}\}}\to\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(M+\Lambda)\,\omega_L\right)\Big/\\
&\quad\left((1+\omega_L)\,\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(1+2\,M+2\,\Lambda)\,\omega_L+\omega_L^2\right)\right),\ \psi_{\{\{a,c\}\},\{\{b\}\}}\to\\
&\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(M+\Lambda)\,\omega_L\right)\Big/\left((1+\omega_L)\,\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(1+2\,M+2\,\Lambda)\,\omega_L+\omega_L^2\right)\right),\\
&\psi_{\{\{b,c\}\},\{\{a\}\}}\to\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(M+\Lambda)\,\omega_L\right)\Big/\\
&\quad\left((1+\omega_L)\,\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(1+2\,M+2\,\Lambda)\,\omega_L+\omega_L^2\right)\right),\ \psi_{\{\{a\},\{b,c\}\},\{\}}\to\\
&\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(1+M+2\,\Lambda)\,\omega_L+\omega_L^2\right)\Big/\left((1+\omega_L)\,\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(1+2\,M+2\,\Lambda)\,\omega_L+\omega_L^2\right)\right),\\
&\psi_{\{\{b\},\{a,c\}\},\{\}}\to\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(1+M+2\,\Lambda)\,\omega_L+\omega_L^2\right)\Big/\\
&\quad\left((1+\omega_L)\,\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(1+2\,M+2\,\Lambda)\,\omega_L+\omega_L^2\right)\right),\ \psi_{\{\{c\},\{a,b\}\},\{\}}\to\\
&\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(1+M+2\,\Lambda)\,\omega_L+\omega_L^2\right)\Big/\left((1+\omega_L)\,\left(M+\Lambda+2\,M\,\Lambda+\Lambda^2+(1+2\,M+2\,\Lambda)\,\omega_L+\omega_L^2\right)\right)\Big\}
\end{aligned}
$$

We have rewritten this in terms of ω_L, which refers to the sum of the ω's for the two lineages involved.

Now we solve for 3 genes in two demes:

```
eqs3 = selectEqns[eqs, {2, 3}];
soln3 = Solve[eqs3, First /@ eqs3]⟦1⟧;
soln3Simp = soln3 /. soln1 /. soln2Simp;
```

As a check, if we set $\omega\to 0$, the GF is always 1, independent of Λ:

```
soln3Simp /. {ω[_] → 0} // Simplify
```

```
{GF[ω, {{}, {{a}, {b}, {c}}}, M, {{x}, {y}}, Λ] → 1,
 GF[ω, {{{a}}, {{b}, {c}}}, M, {{x}, {y}}, Λ] → 1, GF[ω, {{{b}}, {{a}, {c}}}, M, {{x}, {y}}, Λ] → 1,
 GF[ω, {{{c}}, {{a}, {b}}}, M, {{x}, {y}}, Λ] → 1, GF[ω, {{{a}, {b}}, {{c}}}, M, {{x}, {y}}, Λ] → 1,
 GF[ω, {{{a}, {c}}, {{b}}}, M, {{x}, {y}}, Λ] → 1, GF[ω, {{{b}, {c}}, {{a}}}, M, {{x}, {y}}, Λ] → 1,
 GF[ω, {{{a}, {b}, {c}}, {}}, M, {{x}, {y}}, Λ] → 1}
```

■ 1.2 Sumaries for exponentially distributed split times

▫ *The total length of the genealogy*

A relatively simple expression can be obtained for the distribution of total length of the genealogy, $T = t_{\{a\}} + t_{\{b\}} + \ldots$, for a given Λ by setting all the ω to be the same, so that $\psi = E[\exp(-\omega T)]$

```
ss = soln3Simp /. ω[_] → ω /. tidyNotation[ψ] /. Λ{x},{y} → Λ // Simplify
```

$$\begin{aligned}
\Big\{\psi_{\{\},\{\{a\},\{b\},\{c\}\}} \to \Big(&\Lambda^4 + 5\,\Lambda^3\,(1+2\,\omega) + \Lambda^2\,(7+36\,\omega+37\,\omega^2) + 6\,\omega\,(1+6\,\omega+11\,\omega^2+6\,\omega^3) + \\
&\Lambda\,(3+32\,\omega+85\,\omega^2+60\,\omega^3) + M^2\,(3+4\,\Lambda^2+9\,\omega+6\,\omega^2+2\,\Lambda\,(4+7\,\omega)) + \\
&M\,(4\,\Lambda^3+\Lambda^2\,(14+27\,\omega)+\Lambda\,(13+56\,\omega+53\,\omega^2)+3\,(1+7\,\omega+14\,\omega^2+8\,\omega^3))\Big)\Big/ \\
&\Big((1+\omega)\,(1+2\,\omega)\,(M+\Lambda+2\,M\,\Lambda+\Lambda^2+2\,\omega+4\,M\,\omega+4\,\Lambda\,\omega+4\,\omega^2) \\
&(3+\Lambda^2+12\,\omega+9\,\omega^2+\Lambda\,(4+6\,\omega)+M\,(3+2\,\Lambda+6\,\omega))\Big),
\end{aligned}$$

$$\begin{aligned}
\psi_{\{\{a\}\},\{\{b\},\{c\}\}} \to \Big(&M^2\,(3+4\,\Lambda^2+9\,\omega+6\,\omega^2+2\,\Lambda\,(4+7\,\omega)) + \\
&M\,(4\,\Lambda^3+3\,(1+\omega)\,(1+2\,\omega)^2+\Lambda^2\,(14+23\,\omega)+\Lambda\,(13+46\,\omega+37\,\omega^2)) + \\
&\Lambda\,(\Lambda^3+\Lambda^2\,(5+8\,\omega)+\Lambda\,(7+26\,\omega+21\,\omega^2)+3\,(1+6\,\omega+11\,\omega^2+6\,\omega^3))\Big)\Big/ \\
&\Big((1+\omega)\,(1+2\,\omega)\,(M+\Lambda+2\,M\,\Lambda+\Lambda^2+2\,\omega+4\,M\,\omega+4\,\Lambda\,\omega+4\,\omega^2) \\
&(3+\Lambda^2+12\,\omega+9\,\omega^2+\Lambda\,(4+6\,\omega)+M\,(3+2\,\Lambda+6\,\omega))\Big),
\end{aligned}$$

$$\begin{aligned}
\psi_{\{\{b\}\},\{\{a\},\{c\}\}} \to \Big(&M^2\,(3+4\,\Lambda^2+9\,\omega+6\,\omega^2+2\,\Lambda\,(4+7\,\omega)) + \\
&M\,(4\,\Lambda^3+3\,(1+\omega)\,(1+2\,\omega)^2+\Lambda^2\,(14+23\,\omega)+\Lambda\,(13+46\,\omega+37\,\omega^2)) + \\
&\Lambda\,(\Lambda^3+\Lambda^2\,(5+8\,\omega)+\Lambda\,(7+26\,\omega+21\,\omega^2)+3\,(1+6\,\omega+11\,\omega^2+6\,\omega^3))\Big)\Big/ \\
&\Big((1+\omega)\,(1+2\,\omega)\,(M+\Lambda+2\,M\,\Lambda+\Lambda^2+2\,\omega+4\,M\,\omega+4\,\Lambda\,\omega+4\,\omega^2) \\
&(3+\Lambda^2+12\,\omega+9\,\omega^2+\Lambda\,(4+6\,\omega)+M\,(3+2\,\Lambda+6\,\omega))\Big),
\end{aligned}$$

$$\begin{aligned}
\psi_{\{\{c\}\},\{\{a\},\{b\}\}} \to \Big(&M^2\,(3+4\,\Lambda^2+9\,\omega+6\,\omega^2+2\,\Lambda\,(4+7\,\omega)) + \\
&M\,(4\,\Lambda^3+3\,(1+\omega)\,(1+2\,\omega)^2+\Lambda^2\,(14+23\,\omega)+\Lambda\,(13+46\,\omega+37\,\omega^2)) + \\
&\Lambda\,(\Lambda^3+\Lambda^2\,(5+8\,\omega)+\Lambda\,(7+26\,\omega+21\,\omega^2)+3\,(1+6\,\omega+11\,\omega^2+6\,\omega^3))\Big)\Big/ \\
&\Big((1+\omega)\,(1+2\,\omega)\,(M+\Lambda+2\,M\,\Lambda+\Lambda^2+2\,\omega+4\,M\,\omega+4\,\Lambda\,\omega+4\,\omega^2) \\
&(3+\Lambda^2+12\,\omega+9\,\omega^2+\Lambda\,(4+6\,\omega)+M\,(3+2\,\Lambda+6\,\omega))\Big),
\end{aligned}$$

$$\begin{aligned}
\psi_{\{\{a\},\{b\}\},\{\{c\}\}} \to \Big(&M^2\,(3+4\,\Lambda^2+9\,\omega+6\,\omega^2+2\,\Lambda\,(4+7\,\omega)) + \\
&M\,(4\,\Lambda^3+3\,(1+\omega)\,(1+2\,\omega)^2+\Lambda^2\,(14+23\,\omega)+\Lambda\,(13+46\,\omega+37\,\omega^2)) + \\
&\Lambda\,(\Lambda^3+\Lambda^2\,(5+8\,\omega)+\Lambda\,(7+26\,\omega+21\,\omega^2)+3\,(1+6\,\omega+11\,\omega^2+6\,\omega^3))\Big)\Big/ \\
&\Big((1+\omega)\,(1+2\,\omega)\,(M+\Lambda+2\,M\,\Lambda+\Lambda^2+2\,\omega+4\,M\,\omega+4\,\Lambda\,\omega+4\,\omega^2) \\
&(3+\Lambda^2+12\,\omega+9\,\omega^2+\Lambda\,(4+6\,\omega)+M\,(3+2\,\Lambda+6\,\omega))\Big),
\end{aligned}$$

$$\begin{aligned}
\psi_{\{\{a\},\{c\}\},\{\{b\}\}} \to \Big(&M^2\,(3+4\,\Lambda^2+9\,\omega+6\,\omega^2+2\,\Lambda\,(4+7\,\omega)) + \\
&M\,(4\,\Lambda^3+3\,(1+\omega)\,(1+2\,\omega)^2+\Lambda^2\,(14+23\,\omega)+\Lambda\,(13+46\,\omega+37\,\omega^2)) + \\
&\Lambda\,(\Lambda^3+\Lambda^2\,(5+8\,\omega)+\Lambda\,(7+26\,\omega+21\,\omega^2)+3\,(1+6\,\omega+11\,\omega^2+6\,\omega^3))\Big)\Big/ \\
&\Big((1+\omega)\,(1+2\,\omega)\,(M+\Lambda+2\,M\,\Lambda+\Lambda^2+2\,\omega+4\,M\,\omega+4\,\Lambda\,\omega+4\,\omega^2) \\
&(3+\Lambda^2+12\,\omega+9\,\omega^2+\Lambda\,(4+6\,\omega)+M\,(3+2\,\Lambda+6\,\omega))\Big),
\end{aligned}$$

$$\begin{aligned}
\psi_{\{\{b\},\{c\}\},\{\{a\}\}} \to \Big(&M^2\,(3+4\,\Lambda^2+9\,\omega+6\,\omega^2+2\,\Lambda\,(4+7\,\omega)) + \\
&M\,(4\,\Lambda^3+3\,(1+\omega)\,(1+2\,\omega)^2+\Lambda^2\,(14+23\,\omega)+\Lambda\,(13+46\,\omega+37\,\omega^2)) + \\
&\Lambda\,(\Lambda^3+\Lambda^2\,(5+8\,\omega)+\Lambda\,(7+26\,\omega+21\,\omega^2)+3\,(1+6\,\omega+11\,\omega^2+6\,\omega^3))\Big)\Big/ \\
&\Big((1+\omega)\,(1+2\,\omega)\,(M+\Lambda+2\,M\,\Lambda+\Lambda^2+2\,\omega+4\,M\,\omega+4\,\Lambda\,\omega+4\,\omega^2) \\
&(3+\Lambda^2+12\,\omega+9\,\omega^2+\Lambda\,(4+6\,\omega)+M\,(3+2\,\Lambda+6\,\omega))\Big),
\end{aligned}$$

$$\begin{aligned}
\psi_{\{\{a\},\{b\},\{c\}\},\{\}} \to \Big(&\Lambda^4 + 5\,\Lambda^3\,(1+2\,\omega) + \Lambda^2\,(7+36\,\omega+37\,\omega^2) + 6\,\omega\,(1+6\,\omega+11\,\omega^2+6\,\omega^3) + \\
&\Lambda\,(3+32\,\omega+85\,\omega^2+60\,\omega^3) + M^2\,(3+4\,\Lambda^2+9\,\omega+6\,\omega^2+2\,\Lambda\,(4+7\,\omega)) + \\
&M\,(4\,\Lambda^3+\Lambda^2\,(14+27\,\omega)+\Lambda\,(13+56\,\omega+53\,\omega^2)+3\,(1+7\,\omega+14\,\omega^2+8\,\omega^3))\Big)\Big/ \\
&\Big((1+\omega)\,(1+2\,\omega)\,(M+\Lambda+2\,M\,\Lambda+\Lambda^2+2\,\omega+4\,M\,\omega+4\,\Lambda\,\omega+4\,\omega^2) \\
&(3+\Lambda^2+12\,\omega+9\,\omega^2+\Lambda\,(4+6\,\omega)+M\,(3+2\,\Lambda+6\,\omega))\Big)\Big\}
\end{aligned}$$

These are the variables in the more compact notation:

```
vars = GetVars[soln3Simp] /. tidyNotation[ψ]
```

$$\{\psi_{\{\},\{\{a\},\{b\},\{c\}\}},\ \psi_{\{\{a\}\},\{\{b\},\{c\}\}},\ \psi_{\{\{b\}\},\{\{a\},\{c\}\}},\ \psi_{\{\{c\}\},\{\{a\},\{b\}\}},\ \psi_{\{\{a\},\{b\}\},\{\{c\}\}},\ \psi_{\{\{a\},\{c\}\},\{\{b\}\}},\ \psi_{\{\{b\},\{c\}\},\{\{a\}\}},\ \psi_{\{\{a\},\{b\},\{c\}\},\{\}}\}$$

We only have to worry about two kinds of configuration. For three genes in the same deme, Λ makes no difference:

```
Take[vars, 2] /. ss /. M → 0 // Simplify
```

$$\left\{\frac{1}{1+3\,\omega+2\,\omega^2},\ \frac{\Lambda}{(1+\omega)\,(1+2\,\omega)\,(\Lambda+2\,\omega)}\right\}$$

The distribution does depend on M when Λ=0:

```
Take[vars, 2] /. ss /. Λ → 0 // Simplify
```

$$\left\{\left(M+M^2+4\,M\,\omega+2\,\omega\,(1+3\,\omega)\right)\Big/\left(\left(1+M+4\,\omega+2\,M\,\omega+3\,\omega^2\right)\,(M+4\,M\,\omega+2\,\omega\,(1+2\,\omega))\right),\ (M\,(1+M+2\,\omega))\Big/\left(\left(1+M+4\,\omega+2\,M\,\omega+3\,\omega^2\right)\,(M+4\,M\,\omega+2\,\omega\,(1+2\,\omega))\right)\right\}$$

However, the mean length of the genealogy for three genes in the same deme is independent of M for $\Lambda = 0$ - an extension of the result for two genes. This is the full expression for mean length as a function of Λ and M:

```
mnL = (-D[# /. ss, ω] & /@ Take[vars, 2]) /. ω → 0 // Simplify
```

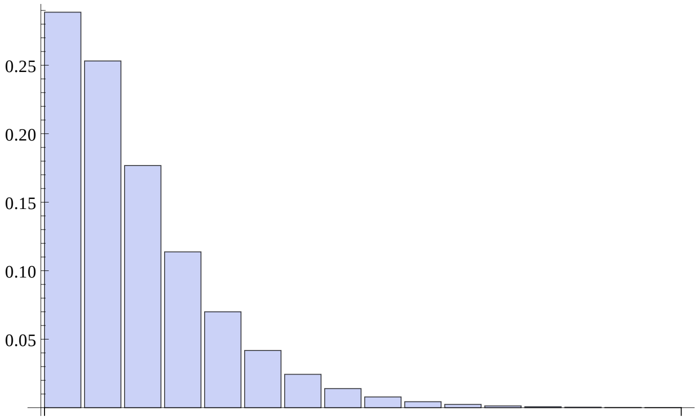
$$\left\{\frac{3\,(1+\Lambda)\,(2\,M+\Lambda)}{M+\Lambda+2\,M\,\Lambda+\Lambda^2},\ \frac{(1+\Lambda)\,(2+6\,M+3\,\Lambda)}{M+\Lambda+2\,M\,\Lambda+\Lambda^2}\right\}$$

□ *# of segregating sites*

The probability that there are X segregating sites in total is $E\left[e^{-\theta\,t/2}\,\frac{(\theta\,t/2)^{X}}{X!}\right]$.

This gives the distribution of # of segregating sites, for M=0.6 , θ=1, Λ=0.7 (three genes in the same deme). Recall that Λ is the rate of splits in scaled time: we are assuming that T is exponentially distributed with mean 1/Λ.

```
cc = CoefficientList[Series[vars⟦1⟧ /. ss /. {Λ → 0.7,
  M → 0.6, ω -> 1/2 - x}, {x, 0, 15}], x] (1/2)^(Range[0, 15]);
BarChart[cc]
{cc, Total[cc]}
```



```
{{0.28869, 0.253114, 0.176776, 0.11377, 0.0700139, 0.0417864, 0.0243667, 0.0139495, 0.00786732,
  0.00438275, 0.00241661, 0.00132101, 0.000716804, 0.000386489, 0.000207243, 0.000110592}, 0.999876}
```

■ 1.3 Sumaries for specific T

□ *The total length of the genealogy*

We can get expressions directly in terms of the split time by taking the ILT wrt Λ:

```
ilT = InverseLaplaceTransform[Λ^-1 Take[vars, 2] /. ss, Λ, T] // Simplify;
```

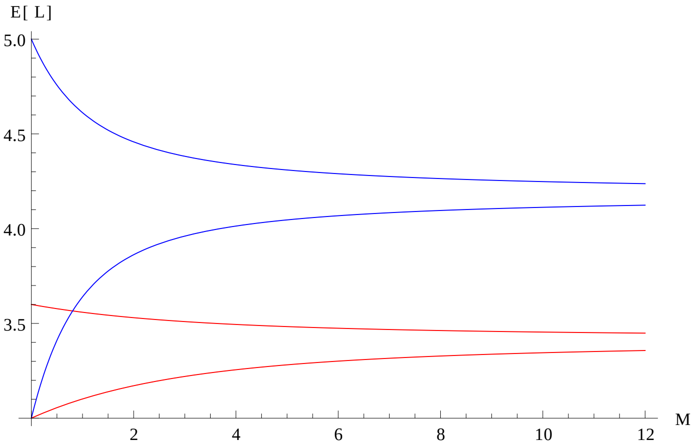
This is the mean length of the genealogy in the IM model with three genes, with sampling configurations {a/b,c} and {a,b,c/∅}:

```
mn = FullSimplify[-D[ilT, ω] /. ω → 0, T > 0]
```

$$\left\{-\left(3\,e^{-\frac{1}{2}\left(1+2M+\sqrt{1+4M^2}\right)T}\left(-1-2M+\sqrt{1+4M^2}-4\,e^{\frac{1}{2}\left(1+2M+\sqrt{1+4M^2}\right)T}\sqrt{1+4M^2}+e^{\sqrt{1+4M^2}\,T}\left(1+2M+\sqrt{1+4M^2}\right)\right)\right)\Big/\left(2\sqrt{1+4M^2}\right),\ \frac{1}{2M\sqrt{1+4M^2}}\,e^{-\frac{1}{2}\left(1+2M+\sqrt{1+4M^2}\right)T}\left(2-2\sqrt{1+4M^2}+4\,e^{\frac{1}{2}\left(1+2M+\sqrt{1+4M^2}\right)T}(1+3M)\sqrt{1+4M^2}+3M\left(1+2M-\sqrt{1+4M^2}\right)-e^{\sqrt{1+4M^2}\,T}\left(2\left(1+\sqrt{1+4M^2}\right)+3M\left(1+2M+\sqrt{1+4M^2}\right)\right)\right)\right\}$$

This shows how the expected length depends on M, for two different diveregence times T=0.3, 1 (red, blue)

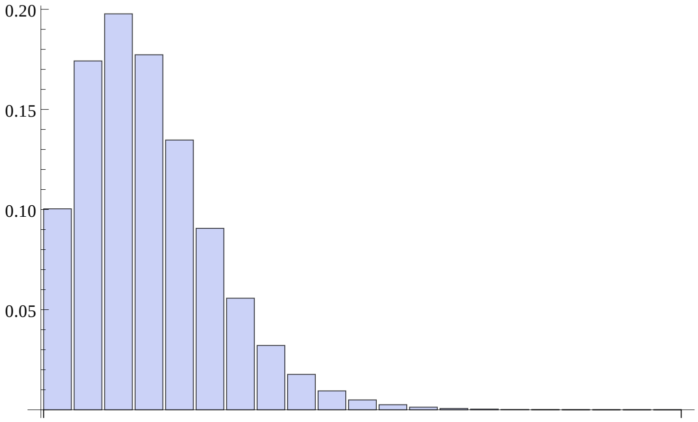
```
Plot[{mn /. {T -> 0.3}, mn /. {T → 1}}, {M, 0, 12},
 PlotStyle → {Red, Blue}, AxesLabel → {"M", "E[L]"}]
```



# of segregating sites

This shows the probability distribution for the total number of segrating sites X for $T = 2$, M=0.6 and θ=1.

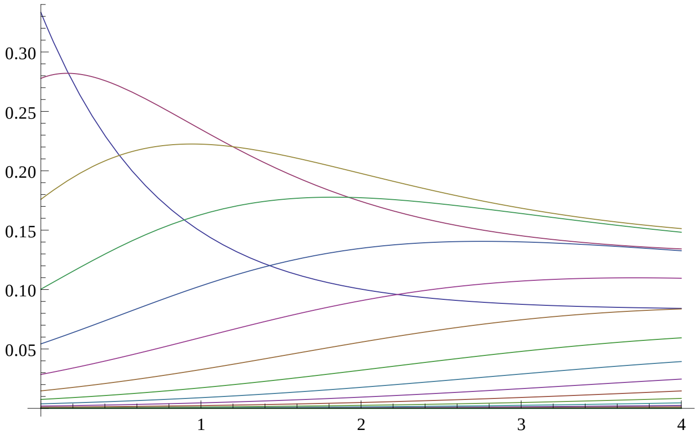
```
ccT = CoefficientList[Series[ilT⟦2⟧ /. {T → 2, M → 0.6, ω → 1/2 - x}, {x, 0, 20}], x] (1/2)^Range[0, 20];
BarChart[ccT]
{ccT, Total[ccT]}
```



```
{{0.100361, 0.17422, 0.197747, 0.177302, 0.13472, 0.0906167, 0.0557276, 0.0321221, 0.0176861,
  0.0094335, 0.00492414, 0.00253334, 0.00129083, 0.000653541, 0.000329491, 0.000165655,
  0.0000831307, 0.0000416664, 0.0000208673, 0.0000104435, 5.23015 × 10^-6}, 0.999995}
```

This shows the probability of 0, 1,…,20 mutations as a function of T; $M = 0.6$, $\theta = 1$.

```
ccT2 = CoefficientList[Series[ilT⟦2⟧ /. {M → 0.6, ω → 1/2 - x}, {x, 0, 20}], x] (1/2)^Range[0, 20];
Plot[ccT2, {T, 0, 4}]
```



▫ *Topological probabilities*

The probability of a particular topology can be found from the LP by taking the limit of the dummy variables corresponding to internal branches incompatible with that topology. For example, to find the probability of a topology {a/b,c} we take the limit of ω_{ab} and ω_{ac} at infinity and set all other ω to zero.

```
{probtopab =
   (Limit[soln3Simp⟦2, 2⟧ /. {ω[{a, c}] → z α, ω[{b, c}] → z, Λ[_] → Λ}, z → ∞] /. ω[_] → 0) //
    Simplify, probtopac =
   (Limit[soln3Simp⟦2, 2⟧ /. {ω[{a, b}] → z α, ω[{b, c}] → z, Λ[_] → Λ}, z → ∞] /. ω[_] → 0) //
    Simplify,
 probtopbc = (Limit[soln3Simp⟦2, 2⟧ /. {ω[{a, b}] → z α, ω[{a, c}] → z, Λ[_] → Λ}, z → ∞] /.
      ω[_] → 0) // Simplify}
```

$$\left\{\frac{2\,M+\Lambda}{3+6\,M+3\,\Lambda},\ \frac{2\,M+\Lambda}{3+6\,M+3\,\Lambda},\ \frac{3+2\,M+\Lambda}{3+6\,M+3\,\Lambda}\right\}$$

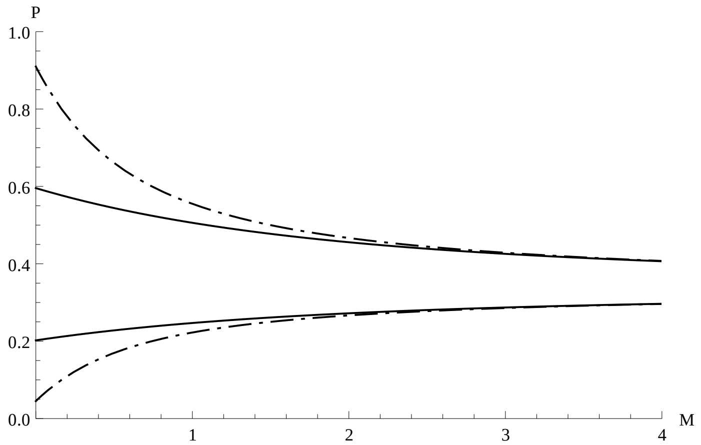
The above sum to one as they should. For a specific time we need to take the ILT of the above and divide by Λ:

```
{probab = InverseLaplaceTransform[probtopab / Λ, Λ, T],
 probac = InverseLaplaceTransform[probtopac / Λ, Λ, T],
 probbc = InverseLaplaceTransform[probtopbc / Λ, Λ, T]}
```

$$\left\{\frac{e^{-(1+2\,M)\,T}}{3\,(1+2\,M)}+\frac{2\,M}{3\,(1+2\,M)},\ \frac{e^{-(1+2\,M)\,T}}{3\,(1+2\,M)}+\frac{2\,M}{3\,(1+2\,M)},\ -\frac{2\,e^{-(1+2\,M)\,T}}{3\,(1+2\,M)}+\frac{3+2\,M}{3\,(1+2\,M)}\right\}$$

This plots topological probabilities for a triplet with sampling configuration {a/b,c} in the symmetric IM model against the scaled migration rate M for two splitting time, T=0.5 (solid lines) and T=2 (dashed lines). The chance of observing an incongruent genealogy {c,{a,b}} or {b{a,c}} (below) increases with M as congruent topologies {a,{b,c}} (above) become less likely.

```
Plot[{{probab, probbc} /. T -> 0.5, {probab, probbc} /. T -> 2},
 {M, 0, 4}, PlotRange -> {{0, 4}, {0, 1}}, AxesLabel → {"M", "P"},
 PlotStyle → {{AbsoluteThickness[1], GrayLevel[0]},
   {AbsoluteThickness[1], GrayLevel[0], AbsoluteDashing[{5, 1, 5}]}}]
```



For samples taken from the same deme, the topologies have the same probability as expected.

```
{(Limit[soln3Simp⟦1, 2⟧ /. {ω[{a, c}] → z α, ω[{b, c}] → z, Λ[_] → Λ}, z → ∞] /. ω[_] → 0) //
  Simplify,
 (Limit[soln3Simp⟦1, 2⟧ /. {ω[{a, b}] → z α, ω[{b, c}] → z, Λ[_] → Λ}, z → ∞] /. ω[_] → 0) //
  Simplify,
 (Limit[soln3Simp⟦1, 2⟧ /. {ω[{a, b}] → z α, ω[{a, c}] → z, Λ[_] → Λ}, z → ∞] /. ω[_] → 0) //
  Simplify}
```

$$\left\{\frac{1}{3},\ \frac{1}{3},\ \frac{1}{3}\right\}$$

▫ *The # of mutations on the internal branch for a given topology*

To find the GF for a particular internal branch conditional on a topology, we take the limit of the ω inconsistent with this topology at infinity and again set ω corresponding to external branches to zero. For branches {a,b} and {b,c} we have:

```
limSolGen2demab = Limit[soln3Simp⟦2, 2⟧ /. {ω[{a, c}] → z α, ω[{b, c}] → z, Λ[_] → Λ}, z → ∞];
limSolGen2dem2ab = limSolGen2demab //. {ω[{a, b}] → ωAB, ω[_] → ω} // Simplify;
limSolGen2dembc = Limit[soln3Simp⟦2, 2⟧ /. {ω[{a, b}] → z α, ω[{a, c}] → z, Λ[_] → Λ}, z → ∞];
limSolGen2dem2bc = limSolGen2dembc //. {ω[{b, c}] → ωBC, ω[_] → ω} // Simplify;
```

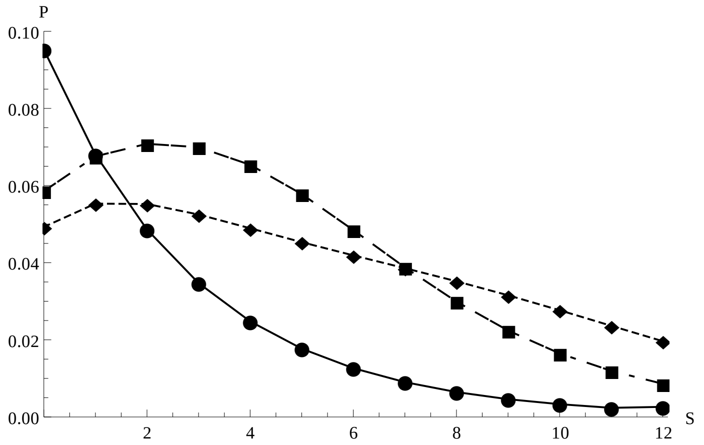
To condition on particular time, we take the ILT at T.

```
iltab = InverseLaplaceTransform[Λ^-1 limSolGen2dem2ab, Λ, T] // Simplify;
iltbc = InverseLaplaceTransform[Λ^-1 limSolGen2dem2bc, Λ, T] // Simplify;
```

```
km = 12;
clab =
  Table[List[Table[i, {i, 0, km}], CoefficientList[Series[iltab /. {ω → 0, ωAB :> 5/2 - yAB, M → 0.8},
        {yAB, 0, km}], yAB] Table[(5/2)^i, {i, 0, km}] // Chop] // Thread, {T, 0, 4, 2}];
clbc = Table[List[Table[i, {i, 0, km}], CoefficientList[Series[iltbc /. {ω -> 0, ωBC :> 5/2 - yBC,
          M → 0.8}, {yBC, 0, km}], yBC] Table[(5/2)^i, {i, 0, km}] // Chop] // Thread, {T, 0, 4, 2}];
```

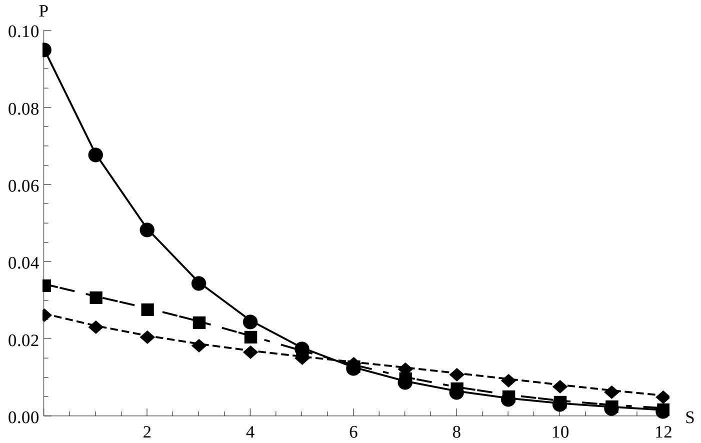
This shows the distribution of the number of mutations on internal the branch {bc} (corresponding to a topology congruent with the sampling configuration) for θ=5, M=0.8 for three different splitting times T = 0 (circles), T = 2 (squares), T = 4 (diamonds):

```
ListPlot[{clbc⟦1⟧, clbc⟦2⟧, clbc⟦3⟧}, PlotRange → {{0, 12.1}, {0, 0.1}}, PlotJoined → True,
 Mesh → All, PlotMarkers → {Automatic, Medium}, MeshStyle → {GrayLevel[0]}, AxesLabel → {"S", "P"},
 PlotStyle → {{AbsoluteThickness[1], GrayLevel[0]}, {AbsoluteThickness[1], GrayLevel[0],
    AbsoluteDashing[{7, 2, 7}]}, {AbsoluteThickness[1], GrayLevel[0], AbsoluteDashing[{3, 3, 3}]}}]
```



This shows the distribution of the number of mutations on internal the branch {a,b} for θ=5, M=0.8 for three different splitting times T=0 (circles), T= 2 (squares), T=4 (diamonds):

```
ListPlot[{clab⟦1⟧, clab⟦2⟧, clab⟦3⟧}, PlotRange → {{0, 12.1}, {0, 0.1}}, PlotJoined → True,
 Mesh → All, PlotMarkers → {Automatic, Medium}, MeshStyle → {GrayLevel[0]}, AxesLabel → {"S", "P"},
 PlotStyle → {{AbsoluteThickness[1], GrayLevel[0]}, {AbsoluteThickness[1], GrayLevel[0],
    AbsoluteDashing[{7, 2, 7}]}, {AbsoluteThickness[1], GrayLevel[0], AbsoluteDashing[{3, 3, 3}]}}]
```



■ 1.4 Full results

□ *Probabilities of mutational configurations for a given topology with exponentially distributed split times*

So far, we have derived results for the total number of segregating sites, by replacing all the branch-specific ω_S by a single ω. Now, we turn to the harder problem of finding the joint probabilities of specific configurations of mutations. This can be done by realising that the GF must be a sum of three terms, each corresponding to a different topology. We obtain the GF for a spcific topology explicitly - both for fixed Λ and for a specific split time, T. When we see an informative mutation (i.e. one shared by two of the leaves), we can just use these expressions to calculate likelihoods. If we only see singletons, we must sum over all three topologies.

Suppose that we observe at least one $\{a, b\}$ mutation. Then, we can delete any terms that depend on $\omega_{\{b,c\}}$ or $\omega_{\{a,c\}}$. The simplest way to do this is to set any terms with these in the denominator to zero. We just do this for three genes with sampling configuration {a/b,c} by taking the second row of soln3Simp:

```
soln3Simp⟦2, 2⟧ /. {Λ[_] → 0.7, M → 0.6} /.
 {       aa__                        aa__                        aa__                        aa__
   ───────────────── :> 0,  ───────────────── :> 0,  ───────────────── :> 0,  ───────────────── :> 0};
   bb__ - ω[{b, c}]         bb__ - ω[{a, c}]         bb__ + ω[{b, c}]         bb__ + ω[{a, c}]
```

This method *fails:* it mistakenly deletes terms that have $\omega[\{a, c\}]$ or $\omega[\{b, c\}]$ in the numerator as well as the denominator. *Mathematica*'s built in Limit[...] function gives the right answer - and without the need to specify Λ or M:

```
limSolGen = Limit[soln3Simp⟦2, 2⟧ /. {ω[{a, c}] → z α, ω[{b, c}] → z, Λ[_] → Λ}, z → ∞];
limSolGen2 = limSolGen //. {ω[{a}] → ωS - ω[{b}] - ω[{c}], ω[{a, b}] → ωAB - ω[{c}]} // Simplify;
```

Necessarily, the remaining terms depend only on $\omega_S = \omega_{\{a\}} + \omega_{\{b\}} + \omega_{\{c\}}$ and on $\omega_{\mathrm{AB}} = \omega_{\{a,b\}} + \omega_{\{c\}}$, which correspond to the number of mutations in the intervals before and after the coalescence of the a and b lineages. The table shows their joint probability distribution obtained by inverting w.r.t. ω_S (top to bottom) and ω_{AB} (left to right). In this example, Λ=0.7, $M = 0.6$ and θ=1.

```
km = 10;
cl =
  CoefficientList[Series[limSolGen2 /. {Λ → 0.7, M → 0.6} /. {ωS :→ 1/2 - yS, ωAB :→ 1/2 - yAB}, {yS, 0, km},
      {yAB, 0, km}], {yS, yAB}] Table[(1/2)^(i+j), {i, 0, km}, {j, 0, km}] // Chop; cl // MatrixForm
```

```
 0.0905292         0.0359305         0.0139067         0.00527662        0.00197083         0.000726866        0.00026533
 0.0314364         0.0125158         0.0048477         0.00183919        0.000686699        0.000253161        0.0000923767
 0.00831566        0.00331877        0.00128632        0.00048805        0.000182197        0.0000671558       0.0000244997
 0.00198247        0.000792698       0.000307428       0.00011666        0.0000435501       0.0000160508       5.8551 × 10^-6
 0.000448601       0.000179639       0.0000697056      0.0000264559      9.87661 × 10^-6    3.64009 × 10^-6    1.32782 × 10^-
 0.0000985425      0.0000395058      0.0000153364      5.8218 × 10^-6    2.17358 × 10^-6    8.01111 × 10^-7    2.9223 × 10^-7
 0.0000212579      8.52985 × 10^-6   3.31257 × 10^-6   1.25769 × 10^-6   4.69602 × 10^-7    1.73089 × 10^-7    6.31414 × 10^-
 4.53323 × 10^-6   1.82022 × 10^-6   7.07099 × 10^-7   2.68507 × 10^-7   1.00265 × 10^-7    3.69582 × 10^-8    1.34826 × 10^-
 9.59436 × 10^-7   3.85443 × 10^-7   1.49769 × 10^-7   5.68794 × 10^-8   2.12413 × 10^-8    7.83007 × 10^-9    2.85656 × 10^-
 2.02042 × 10^-7   8.12006 × 10^-8   3.15578 × 10^-8   1.19863 × 10^-8   4.47651 × 10^-9    1.65023 × 10^-9    6.02058 × 10^-
 4.24028 × 10^-8   1.70469 × 10^-8   6.62609 × 10^-9   2.51695 × 10^-9   9.40055 × 10^-10   3.46558 × 10^-10   1.2644 × 10^-1
```

Note that the first column represents the probability that there is no {*a* , *b* } mutation - contrary to the assumption. It should be deleted. If it is included, then the total is equal to the probability of an ab topology, as expected.

```
{Total[First /@ cl], probtopac /. {Λ → 0.7, M -> 0.6}, Total[Total[cl]]}
```

```
{0.132838, 0.218391, 0.218387}
```

▫ *Probabilities of mutational configurations for a given topology with a specific T*

Now, we try doing the same for a specific *time* rather than a specific Λ. That requires that we keep the expressions as functions of Λ and then take the inverse Laplace transform. The expression is ugly, but not too large. Note that the full GF is obtained just by summing the two other terms for the two other possible topologies:

```
ilt = InverseLaplaceTransform[Λ^-1 limSolGen2, Λ, T] // Simplify;
ilt /. {ωS → ω_S, ωAB → ω_AB} //.
  {Sqrt[1 + 4 M + 16 M^2] → α, 1 + 8 M + Sqrt[1 + 64 M^2] + 2 ω_AB → 2 β, Sqrt[1 + 64 M^2] → γ};
```

```
km = 10;
cl = CoefficientList[Series[ilt /. {ωS :> 1/2 - yS, ωAB :> 1/2 - yAB, M → 0.6, T → 2}, {yS, 0, km},
     {yAB, 0, km}], {yS, yAB}] Table[(1/2)^(i+j), {i, 0, km}, {j, 0, km}] // Chop; cl // TableForm
```

```
0.0577467        0.0305081        0.0134862        0.00524047       0.00188228       0.000648139       0.00021878
0.0284365        0.0134435        0.00544754       0.00200202       0.00069707       0.000236416       0.0000792902
0.0102248        0.00434233       0.00163887       0.000578277      0.000197283      0.0000663165      0.0000221659
0.00302825       0.00118388       0.000425197      0.000146215      0.0000493046     0.0000164978      5.50554×10^-6
0.000764908      0.000282339      0.0000982776     0.0000333009     0.000011162      3.72691×10^-6     1.24287×10^-6
0.000168863      0.000060004      0.0000204992     6.89109×10^-6    2.30303×10^-6    7.68233×10^-7     2.56125×10^-7
0.0000333213     0.0000115524     3.90397×10^-6    1.30693×10^-6    4.36171×10^-7    1.45435×10^-7     4.84831×10^-8
6.0019×10^-6     2.0484×10^-6     6.87933×10^-7    2.29804×10^-7    7.66445×10^-8    2.55489×10^-8     8.52027×10^-9
1.00605×10^-6    3.39987×10^-7    1.13782×10^-7    3.79666×10^-8    1.266×10^-8      4.21899×10^-9     1.43155×10^-9
1.59682×10^-7    5.36385×10^-8    1.79163×10^-8    5.97454×10^-9    1.99449×10^-9    6.56397×10^-10    2.41574×10^-10
2.43645×10^-8    8.15498×10^-9    2.72118×10^-9    9.05399×10^-10   3.0696×10^-10    0                 1.59591×10^-10
```

Again, the first column represents the probability that there is no $\{a, \; b\}$ mutation - contrary to the assumption. If it is included, then the total is the probability of an ab topology as expected.

```
{Total[Total[cl]], probab /. {M → 0.6, T → 2}}
{0.183676, 0.183678}
```

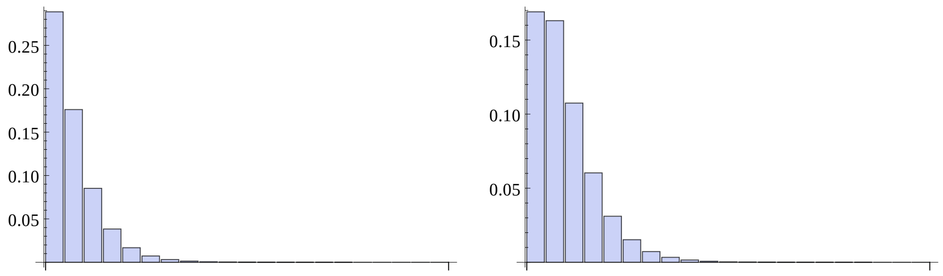
▫ *The # of singletons when there are no informative mutations*

This shows the distribution of the # of singletons for triplets with sampling configurations {a,b,c/∅} (i.e. all samples from the same deme)(left plot) and {a/b,c} (right plot). We assume that there are no mutations on internal branches (i.e., ancestral to two genes) by setting $\omega_{\{\_,\_\}}$ to the scaled mutation rate $\theta/2$. We have chosen specific values θ=1, M=0.6, Λ=0.7.

```
(singletons = soln3Simp /. tidyNotation[ψ] /. {ω{_,_} → 1/2, Λ__ → 0.7, M → 0.6});
{sing1 =
   CoefficientList[Series[singletons⟦1, 2⟧ /. ω{i_} :> 1/2 - y, {y, 0, 20}], y] (1/2)^Range[0,20],
  sing2 = CoefficientList[Series[singletons⟦2, 2⟧ /. ω{i_} :> 1/2 - y, {y, 0, 20}], y]
    (1/2)^Range[0,20]} // TableForm
```

```
0.28869      0.176037     0.0852381     0.0383407     0.016761      0.00725279     0.00313281     0.00135589
0.169064     0.163077     0.107474      0.0603337     0.0310932     0.0152276      0.00721991     0.00335067
```

```
Show[GraphicsGrid[{{BarChart[sing1], BarChart[sing2]}}]]
```



The probability that there will be no informative mutations, for genes in the same vs in different demes is the sum of the tables above, but is obtained more directly by setting ω to zero:

```
{singletons⟦1, 2⟧, singletons⟦2, 2⟧} /. ω{_} → 0
```

```
{0.617854, 0.55963}
```

2. Test on real and simulated data

- **Set up**

The above solutions can be used to compute the joint Log likelihood (LogL) of IM model parameters from very large numbers of loci. If we assume for the moment that loci have the same mutation rate, this requires tabulating the LogL for all observed mutational configurations, multiplying by the number of loci with each configuration. For a triplet sample with sampling configuration {a/b,c} there are three topology classes; loci may be topologically congruent (those with a bc mutation), incongruent (those with an ab or ac mutation) or uninformative. Note that we are assuming outgroup rooting such that each locus can be assigned to the three classes unambiguously (ways of dealing with finite sites mutations are discussed in the last section).

For any rooted topology, there are 3 types of mutations. For example, assuming topology {a/b,c}, we need to distinguish mutations on the internal branch (k_{bc}), those on the shorter external branches k_{ex} (since branches connected to b and c have the same length these can be lumped) and mutations on the longer external branch k_a. However, as shown before, we have the constraint $t_a = t_{b\ c} + t_b = t_{b\ c} + t_c$ and thus the GF is a function only of $\omega_{b\ c} - \omega_a$ and of $\omega_{\text{ex}} - \omega_a$, which correspond to the number of mutations in the two coalescence intervals. The joint probability of the three types of observable mutations $P[k_{\text{bc}}, k_{\text{ex}}, k_a]$ can be found by summing over all possible ways these can be partitioned amongst the two coalescent intervals:

$$P[k_{\text{bc}}, k_{\text{ex}}, k_a] = \sum_{j=0}^{k_a} \binom{k_{\text{ex}} + k_a - j}{k_a - j} \left(\frac{1}{3}\right)^{k_a - j} \left(\frac{2}{3}\right)^{k_{\text{ex}}} \binom{k_{\text{bc}} + j}{j} \left(\frac{1}{2}\right)^{k_{\text{bc}} + j} P[k_{b\ c} + j, k_{\text{ex}} + k_a - j]$$

We need to evaluate the GF for the number of mutations in each coalescence interval for all 3 topology classes. For the IM model with symmetric migration we have (note that for the topologically uninformative loci, we are only using the distribution of the total number of singleton rather than their full, joint distribution here) :

```
limSolGenCON = Limit[soln3Simp⟦2, 2⟧ /. {ω[{a, b}] → z α, ω[{a, c}] → z, Λ[_] → Λ}, z → ∞] //.
    {ω[{b}] → θS - ω[{a}] - ω[{c}], ω[{b, c}] → θBC - ω[{a}]} // Simplify;

limSolGenINCON = Limit[soln3Simp⟦2, 2⟧ /. {ω[{a, c}] → z α, ω[{b, c}] → z, Λ[_] → Λ}, z → ∞] //.
    {ω[{a}] → θS - ω[{b}] - ω[{c}], ω[{a, b}] → θAB - ω[{c}]} // Simplify;

limSolGenNOTOP = soln3Simp⟦2, 2⟧ /. {ω[{a, b}] → θ/2, ω[{b, c}] → θ/2,
     ω[{a, c}] → θ/2, Λ[_] → Λ, ω[{a}] → ωS, ω[{c}] → ωS, ω[{b}] → ωS} // Simplify;

ilt2typesCON = InverseLaplaceTransform[Λ^-1 limSolGenCON, Λ, T] // Simplify;

ilt2typesINCON = InverseLaplaceTransform[Λ^-1 limSolGenINCON, Λ, T] // Simplify;

ilt2typesNOTOP = InverseLaplaceTransform[Λ^-1 limSolGenNOTOP, Λ, T] // Simplify;
```

■ IM with asymmetric migration

Allowing for migration in one direction only greatly simplifies the problem. For a sample of three genes ((a) sampled one and (b) and (c) from the other), we can write down the GF by hand. Assuming that only lineage (a) can have been affected by migration (for migration in the reverse direction see section below) we have 6 equations:

$$
\begin{aligned}
\texttt{asym} = \Big\{ & \psi[\{\}, \{\{a\}, \{b, c\}\}] == \frac{1}{1 + \omega[\{a\}] + \omega[\{b, c\}]}, \\
& \psi[\{\}, \{\{b\}, \{a, c\}\}] == \frac{1}{1 + \omega[\{b\}] + \omega[\{a, c\}]}, \\
& \psi[\{\}, \{\{c\}, \{a, b\}\}] == \frac{1}{1 + \omega[\{c\}] + \omega[\{a, b\}]}, \\
& \psi[\{\}, \{\{a\}, \{b\}, \{c\}\}] == \frac{1}{3 + \omega[\{a\}] + \omega[\{b\}] + \omega[\{c\}]} \\
& \quad (\psi[\{\}, \{\{a\}, \{b, c\}\}] + \psi[\{\}, \{\{b\}, \{a, c\}\}] + \psi[\{\}, \{\{c\}, \{a, b\}\}]), \\
& \psi[\{\{a\}\}, \{\{b, c\}\}] == \frac{1}{\Lambda + (M / 2) + \omega[\{a\}] + \omega[\{b, c\}]} (\Lambda + (M / 2))\ \psi[\{\}, \{\{a\}, \{b, c\}\}], \\
& \psi[\{\{a\}\}, \{\{b\}, \{c\}\}] == \frac{1}{\Lambda + 1 + (M / 2) + \omega[\{a\}] + \omega[\{b\}] + \omega[\{c\}]} \\
& \quad ((\Lambda + (M / 2))\ \psi[\{\}, \{\{a\}, \{b\}, \{c\}\}] + \psi[\{\{a\}\}, \{\{b, c\}\}]) \Big\};
\end{aligned}
$$

▫ *GF conditional on topology*

Solving the above gives the GF for a sample (a, (b,c)):

```
asymGF = (Solve[asym, First /@ asym])⟦1, -1, 2⟧ // Simplify
```

```
((M + 2 Λ)
   ((2 + ω[{b}] + ω[{c}] + ω[{a, b}] + ω[{a, c}]) / ((1 + ω[{c}] + ω[{a, b}]) (1 + ω[{b}] + ω[{a, c}])) +
     (6 + M + 2 Λ + 4 ω[{a}] + 2 ω[{b}] + 2 ω[{c}] + 2 ω[{b, c}]) /
      ((1 + ω[{a}] + ω[{b, c}]) (M + 2 Λ + 2 ω[{a}] + 2 ω[{b, c}])))) /
 ((3 + ω[{a}] + ω[{b}] + ω[{c}]) (2 + M + 2 Λ + 2 ω[{a}] + 2 ω[{b}] + 2 ω[{c}]))
```

In this case we can invert wrt Λ to find the GF for a discrete splitting time T. The expression is complex but not vast...

```
asymGF2 = InverseLaplaceTransform[Λ^-1 asymGF, Λ, T] // Simplify
```

```
((2 ⅇ^(-1/2 T (M+2 ω[{a}]+2 ω[{b,c}])) (3 + ω[{a}] + ω[{b}] + ω[{c}]) (ω[{a}] + ω[{b, c}]))/
    ((1 + ω[{b}] + ω[{c}] - ω[{b, c}]) (M + 2 ω[{a}] + 2 ω[{b, c}])) +
   (2 ⅇ^(-1/2 T (2+M+2 ω[{a}]+2 ω[{b}]+2 ω[{c}])) (1 + ω[{a}] + ω[{b}] + ω[{c}])
      (ω[{c}] + ω[{c}]^2 - ω[{a, b}] + ω[{c}] ω[{a, b}] - ω[{a, c}] - ω[{c}] ω[{a, c}] - 2 ω[{a, b}]
         ω[{a, c}] - ω[{b, c}] + ω[{c}] ω[{b, c}] + ω[{c}]^2 ω[{b, c}] - ω[{a, b}] ω[{b, c}] +
        ω[{c}] ω[{a, b}] ω[{b, c}] - ω[{a, c}] ω[{b, c}] - ω[{a, b}] ω[{a, c}] ω[{b, c}] -
        2 ω[{b, c}]^2 - ω[{c}] ω[{b, c}]^2 - ω[{a, b}] ω[{b, c}]^2 - ω[{a, c}] ω[{b, c}]^2 +
        ω[{b}]^2 (1 + ω[{b, c}]) + ω[{a}] (1 + ω[{b}]^2 + ω[{c}]^2 - ω[{a, b}] ω[{a, c}] +
           ω[{c}] (2 + ω[{a, b}] - ω[{b, c}]) + ω[{b}] (2 + ω[{c}] + ω[{a, c}] - ω[{b, c}]) -
           2 ω[{b, c}] - ω[{a, b}] ω[{b, c}] - ω[{a, c}] ω[{b, c}]) + ω[{b}]
         (1 - ω[{a, b}] + ω[{b, c}] + ω[{c}] ω[{b, c}] - ω[{b, c}]^2 + ω[{a, c}] (1 + ω[{b, c}]))))/
    ((2 + M + 2 ω[{a}] + 2 ω[{b}] + 2 ω[{c}]) (1 + ω[{c}] + ω[{a, b}])
      (1 + ω[{b}] + ω[{a, c}]) (1 + ω[{b}] + ω[{c}] - ω[{b, c}])) +
   (M (6 + 3 M + 8 ω[{c}] + 2 M ω[{c}] + 2 ω[{c}]^2 + 6 ω[{a, b}] + 2 M ω[{a, b}] +
        2 ω[{c}] ω[{a, b}] + 2 ω[{b}]^2 (1 + ω[{c}] + ω[{a, b}]) + 6 ω[{a, c}] +
        2 M ω[{a, c}] + 8 ω[{c}] ω[{a, c}] + M ω[{c}] ω[{a, c}] + 2 ω[{c}]^2 ω[{a, c}] +
        6 ω[{a, b}] ω[{a, c}] + M ω[{a, b}] ω[{a, c}] + 2 ω[{c}] ω[{a, b}] ω[{a, c}] +
        2 ω[{a}]^2 (2 + ω[{b}] + ω[{c}] + ω[{a, b}] + ω[{a, c}]) + 6 ω[{b, c}] + 2 M ω[{b, c}] +
        4 ω[{c}] ω[{b, c}] + M ω[{c}] ω[{b, c}] + 4 ω[{a, b}] ω[{b, c}] + M ω[{a, b}] ω[{b, c}] +
        4 ω[{a, c}] ω[{b, c}] + M ω[{a, c}] ω[{b, c}] + 2 ω[{c}] ω[{a, c}] ω[{b, c}] +
        2 ω[{a, b}] ω[{a, c}] ω[{b, c}] + 4 ω[{b, c}]^2 + 2 ω[{c}] ω[{b, c}]^2 +
        2 ω[{a, b}] ω[{b, c}]^2 + 2 ω[{a, c}] ω[{b, c}]^2 + ω[{b}] (8 + 2 M + 2 ω[{c}]^2 + 2 ω[{a, c}] +
           4 ω[{b, c}] + M ω[{b, c}] + 2 ω[{b, c}]^2 + ω[{a, b}] (8 + M + 2 ω[{a, c}] + 2 ω[{b, c}]) +
           ω[{c}] (10 + M + 2 ω[{a, b}] + 2 ω[{a, c}] + 2 ω[{b, c}])) + ω[{a}]
         (8 + 2 M + 6 ω[{a, b}] + M ω[{a, b}] + 6 ω[{a, c}] + M ω[{a, c}] + 4 ω[{a, b}] ω[{a, c}] +
           8 ω[{b, c}] + 4 ω[{a, b}] ω[{b, c}] + 4 ω[{a, c}] ω[{b, c}] + ω[{b}] (6 + M + 4 ω[{c}] +
              4 ω[{a, b}] + 4 ω[{b, c}]) + ω[{c}] (6 + M + 4 ω[{a, c}] + 4 ω[{b, c}]))))/
    ((2 + M + 2 ω[{a}] + 2 ω[{b}] + 2 ω[{c}]) (1 + ω[{c}] + ω[{a, b}]) (1 + ω[{b}] + ω[{a, c}])
      (M + 2 ω[{a}] + 2 ω[{b, c}])))/
 ((3 + ω[{a}] + ω[{b}] + ω[{c}]) (1 + ω[{a}] + ω[{b, c}]))
```

The GF conditional on a particular topology (congruent or incongruent) can be found by taking the limits as before. The GF only depends on the θBC and θS corresponding to the two coalescence intervals:

```
limSolGenCON2 = Limit[asymGF /. {ω[{a, b}] → z α, ω[{a, c}] → z}, z → ∞] //.
   {ω[{b}] → θS - ω[{a}] - ω[{c}], ω[{b, c}] → θBC - ω[{a}]} // Simplify
```

```
((M + 2 Λ) (M + 2 (3 + θBC + θS + Λ))) / ((1 + θBC) (3 + θS) (M + 2 (θBC + Λ)) (M + 2 (1 + θS + Λ)))
```

```
limSolGenINCON2 = Limit[asymGF /. {ω[{a, c}] → z α, ω[{b, c}] → z}, z → ∞] //.
   {ω[{a}] → θS - ω[{b}] - ω[{c}], ω[{a, b}] → θAB - ω[{c}]} // Simplify
```

```
               M + 2 Λ
───────────────────────────────────────
(1 + θAB) (3 + θS) (M + 2 (1 + θS + Λ))
```

Inverting the above wrt Λ gives:

```
ilt2typesCON2 = InverseLaplaceTransform[Λ^-1 limSolGenCON2, Λ, T] // Simplify
```

$$\left(-\frac{2\, e^{-\frac{1}{2} T (M+2\, \theta BC)}\, \theta BC\, (3+\theta S)}{(M+2\, \theta BC)\, (-1+\theta BC-\theta S)} - \frac{2\, e^{-\frac{1}{2} T (2+M+2\, \theta S)}\, (2+\theta BC)\, (1+\theta S)}{(1-\theta BC+\theta S)\, (2+M+2\, \theta S)} + \frac{M\, (M+2\, (3+\theta BC+\theta S))}{(M+2\, \theta BC)\, (2+M+2\, \theta S)}\right) \Big/ ((1+\theta BC)\, (3+\theta S))$$

```
ilt2typesINCON2 = InverseLaplaceTransform[Λ^-1 limSolGenINCON2, Λ, T] // Simplify
```

$$\frac{M+2\, e^{-\frac{1}{2} T (2+M+2\, \theta S)}\, (1+\theta S)}{(1+\theta AB)\, (3+\theta S)\, (2+M+2\, \theta S)}$$

Check

Setting all ω to zero the GF must sum to one:

```
{asymGF /. {ω[_] -> 0}, asymGF2 /. {ω[_] → 0}} // Simplify
```

```
{1, 1}
```

Topological probabilities sum to one as they should:

```
{topcon = ilt2typesCON2 /. {θS → 0, θBC → 0},
  topincon = ilt2typesINCON2 /. {θS → 0, θAB → 0}} // Simplify
```

$$\left\{\frac{6-4\, e^{-\frac{1}{2} (2+M)\, T}+M}{3\, (2+M)}, \frac{2\, e^{-\frac{1}{2} (2+M)\, T}+M}{3\, (2+M)}\right\}$$

```
topcon + 2 topincon // FullSimplify
```

```
1
```

GF for topologically uninformative blocks

To obtain the GF for topologically uninformative blocks we need to sum over all three possible topologies

```
limSolGenNOTOPbc = Limit[asymGF /. {ω[{a, b}] → z α, ω[{a, c}] → z, ω[{b, c}] → θ/2}, z → ∞] /.
   {ω[{c}] → ωsh - ω[{b}], ω[{a}] → ωa} // Simplify;
limSolGenNOTOPab = Limit[asymGF /. {ω[{a, c}] → z α, ω[{b, c}] → z, ω[{a, b}] → θ/2}, z → ∞] /.
   {ω[{a}] → ωsh - ω[{b}], ω[{c}] → ωc} // Simplify;
limSolGenNOTOPac = Limit[asymGF /. {ω[{a, b}] → z α, ω[{b, c}] → z, ω[{a, c}] → θ/2}, z → ∞] /.
   {ω[{a}] → ωsh - ω[{c}], ω[{b}] → ωb} // Simplify;
```

To GF for discrete splitting times are:

```
ilt2typesNOTOPbc = InverseLaplaceTransform[Λ^-1 limSolGenNOTOPbc, Λ, T] // Simplify;
ilt2typesNOTOPab = InverseLaplaceTransform[Λ^-1 limSolGenNOTOPab, Λ, T] // Simplify;
ilt2typesNOTOPac = InverseLaplaceTransform[Λ^-1 limSolGenNOTOPac, Λ, T] // Simplify;
```

GF for Total S

The GF for the total number of mutations S is found by setting all ω[_] to be the same:

```
GfS = InverseLaplaceTransform[Λ^-1 (asymGF /. {ω[_] → ω} // Simplify), Λ, T] // Simplify
```

$$\frac{M + 4\, e^{-\frac{1}{2} T (M+4\omega)}\, \omega}{(1+\omega)\,(1+2\,\omega)\,(M+4\,\omega)}$$

This tabulates the pdfF of S :

```
test = probSasym[0.127, 4.2, 0.5, 12]
```

```
{0.0601006, 0.0853815, 0.0888486, 0.08368, 0.075961, 0.0679905,
 0.0605291, 0.0537757, 0.0477385, 0.0423665, 0.0375948, 0.0333591, 0.0296002}
```

```
0.0635 * 2
```

```
0.127
```

```
test = probSasym[0.0635, 4.2, 0.5, 12]
```

```
{0.104485, 0.197259, 0.228484, 0.194915, 0.132674, 0.0759036, 0.0379126,
 0.0170273, 0.00704105, 0.00273326, 0.00101202, 0.000362034, 0.000126404}
```

Which again must sum to one:

```
test // Total
```

```
0.999991
```

▫ *Pairwise GF*

The GF for the pairwise coalescence times for the asymmetric case is:

$$\psi\,[\,\mathrm{diff}\,] = \frac{1}{\Lambda + M/2 + \omega}\;(\,M/2 + \Lambda)\;\frac{1}{(1+\omega)}\;;$$

```
InverseLaplaceTransform[ψ [diff] / Λ, Λ, T]
```

$$\frac{\frac{M}{M+2\,\omega} + \frac{2\, e^{-\frac{1}{2} T (M+2\omega)}\,\omega}{M+2\,\omega}}{1+\omega}$$

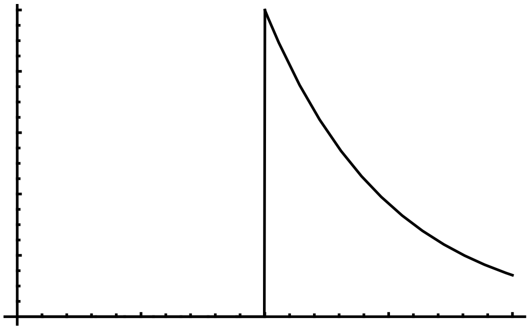
```
InverseLaplaceTransform[ψ [diff] / Λ, Λ, T] /. {M → 0}
```

$$\frac{e^{-T\,\omega}}{1+\omega}$$

```
PDFcoal = InverseLaplaceTransform[InverseLaplaceTransform[ψ [diff] / Λ, Λ, T], ω, t]
```

$$\frac{1}{-2+M}\, e^{-\frac{1}{2}(2+M)\,t}\left(\left(-e^{t} + e^{\frac{M\,t}{2}}\right) M + \left(-2\, e^{\frac{1}{2} M\,(t-T)+T} + e^{t}\, M\right) \mathrm{HeavisideTheta}[t-T]\right)$$

```
Plot[PDFcoal /. {M → 0, T → 2}, {t, 0, 4}] // N
```



```
- Graphics -
```

▫ ***Migration in the reverse direction...***

With migration in the reverse direction we have 10 equations:

$$\mathbf{asymREV} = \Big\{\psi[\{\{a\}, \{b, c\}\}, \{\}] == \frac{1}{1 + \omega[\{a\}] + \omega[\{b, c\}]},$$

$$\psi[\{\{b\}, \{a, c\}\}, \{\}] == \frac{1}{1 + \omega[\{b\}] + \omega[\{a, c\}]},$$

$$\psi[\{\{c\}, \{a, b\}\}, \{\}] == \frac{1}{1 + \omega[\{c\}] + \omega[\{a, b\}]},$$

$$\psi[\{\{a, b\}\}, \{\{c\}\}] == \frac{1}{\Lambda + M/2 + \omega[\{a, b\}] + \omega[\{c\}]} (\Lambda + M/2)\, \psi[\{\{c\}, \{a, b\}\}, \{\}],$$

$$\psi[\{\{a, c\}\}, \{\{b\}\}] == \frac{1}{\Lambda + M/2 + \omega[\{a, c\}] + \omega[\{b\}]} (\Lambda + M/2)\, \psi[\{\{b\}, \{a, c\}\}, \{\}],$$

$$\psi[\{\{a\}, \{b\}\}, \{\{c\}\}] == \frac{1}{\Lambda + 1 + M/2 + \omega[\{c\}] + \omega[\{b\}] + \omega[\{a\}]} ((\Lambda + M/2)\, \psi[\{\{a\}, \{b\}, \{c\}\}, \{\}] + \psi[\{\{a, b\}\}, \{\{c\}\}]),$$

$$\psi[\{\{a\}, \{c\}\}, \{\{b\}\}] == \frac{1}{\Lambda + 1 + M/2 + \omega[\{c\}] + \omega[\{b\}] + \omega[\{a\}]} ((\Lambda + M/2)\, \psi[\{\{a\}, \{b\}, \{c\}\}, \{\}] + \psi[\{\{a, c\}\}, \{\{b\}\}]),$$

$$\psi[\{\{a\}, \{b\}, \{c\}\}, \{\}] == \frac{1}{3 + \omega[\{a\}] + \omega[\{b\}] + \omega[\{c\}]} (\psi[\{\{a\}, \{b, c\}\}, \{\}] + \psi[\{\{b\}, \{a, c\}\}, \{\}] + \psi[\{\{c\}, \{a, b\}\}, \{\}]),$$

$$\psi[\{\{a\}\}, \{\{b, c\}\}] == \frac{1}{\Lambda + (M/2) + \omega[\{a\}] + \omega[\{b, c\}]} (\Lambda + (M/2))\, \psi[\{\{a\}, \{b, c\}\}, \{\}],$$

$$\psi[\{\{a\}\}, \{\{b\}, \{c\}\}] == \frac{1}{\Lambda + 1 + M + \omega[\{a\}] + \omega[\{b\}] + \omega[\{c\}]} (\Lambda\, \psi[\{\{a\}, \{b\}, \{c\}\}, \{\}] + \psi[\{\{a\}\}, \{\{b, c\}\}] + M/2\, \psi[\{\{a\}, \{b\}\}, \{\{c\}\}] + M/2\, \psi[\{\{a\}, \{c\}\}, \{\{b\}\}])\Big\};$$

The GF is:

```
asymGFREV = (Solve[asymREV, First /@ asymREV])⟦1, -1, 2⟧ // Simplify
```

$$\Bigg(\left(\Lambda+\frac{M\left(\frac{M}{2}+\Lambda\right)}{1+\frac{M}{2}+\Lambda+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}]}\right)\Big/\left((3+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}])\,(1+\omega[\{c\}]+\omega[\{a,b\}])\right)+$$

$$(M\,(M+2\,\Lambda))/((2+M+2\,\Lambda+2\,\omega[\{a\}]+2\,\omega[\{b\}]+2\,\omega[\{c\}])\,(1+\omega[\{c\}]+\omega[\{a,b\}])\,(M+2\,\Lambda+2\,\omega[\{c\}]+2\,\omega[\{a,b\}]))+$$

$$\left(\frac{\Lambda+\frac{M\left(\frac{M}{2}+\Lambda\right)}{1+\frac{M}{2}+\Lambda+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}]}}{3+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}]}+(M\,(M+2\,\Lambda))/((2+M+2\,\Lambda+2\,\omega[\{a\}]+2\,\omega[\{b\}]+2\,\omega[\{c\}])\,(M+2\,\Lambda+2\,\omega[\{b\}]+2\,\omega[\{a,c\}]))\right)\Big/(1+\omega[\{b\}]+\omega[\{a,c\}])+$$

$$\left(\frac{\Lambda+\frac{M\left(\frac{M}{2}+\Lambda\right)}{1+\frac{M}{2}+\Lambda+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}]}}{3+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}]}+\frac{M+2\,\Lambda}{M+2\,\Lambda+2\,\omega[\{a\}]+2\,\omega[\{b,c\}]}\right)\Big/(1+\omega[\{a\}]+\omega[\{b,c\}])\Bigg)\Big/$$

$$(1+M+\Lambda+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}])$$

There are higher order terms in M which are not present with the reverse simpler sampling scheme (a single individual from the receiving population).

```
limSolGenCON2REV = Limit[asymGFREV /. {ω[{a, b}] → z α, ω[{a, c}] → z}, z → ∞] //.
   {ω[{b}] → θS - ω[{a}] - ω[{c}], ω[{b, c}] → θBC - ω[{a}]} // Simplify
```

$$\left(M^3+4\,\Lambda\,(1+\theta S+\Lambda)\,(3+\theta BC+\theta S+\Lambda)+M^2\,(3+2\,\theta BC+\theta S+5\,\Lambda)+2\,M\left(3+\theta S^2+7\,\Lambda+3\,\theta BC\,\Lambda+4\,\Lambda^2+\theta S\,(4+3\,\Lambda)\right)\right)\Big/$$
$$((1+\theta BC)\,(3+\theta S)\,(1+M+\theta S+\Lambda)\,(M+2\,(\theta BC+\Lambda))\,(M+2\,(1+\theta S+\Lambda)))$$

```
limSolGenINCON2REV = Limit[asymGFREV /. {ω[{a, c}] → z α, ω[{b, c}] → z}, z → ∞] //.
   {ω[{a}] → θS - ω[{b}] - ω[{c}], ω[{a, b}] → θAB - ω[{c}]} // Simplify
```

$$\left(M^3+4\,\Lambda\,(\theta AB+\Lambda)\,(1+\theta S+\Lambda)+2\,M\,\Lambda\,(4+3\,\theta AB+2\,\theta S+4\,\Lambda)+M^2\,(3+2\,\theta AB+\theta S+5\,\Lambda)\right)\Big/$$
$$((1+\theta AB)\,(3+\theta S)\,(1+M+\theta S+\Lambda)\,(M+2\,(\theta AB+\Lambda))\,(M+2\,(1+\theta S+\Lambda)))$$

Inverting the above wrt Λ gives:

```
ilt2typesCON2REV = InverseLaplaceTransform[Λ^-1 limSolGenCON2REV, Λ, T] // Simplify
```

$$\Bigg(\frac{4\,e^{-\frac{1}{2}T\,(2+M+2\,\theta S)}\,(1+\theta S)}{2+M+2\,\theta S}+\frac{4\,e^{-\frac{1}{2}T\,(M+2\,\theta BC)}\,\theta BC\,(3+\theta S)}{(M+2\,\theta BC)\,(2+M-2\,\theta BC+2\,\theta S)}-$$
$$\left(2\,e^{-T\,(1+M+\theta S)}\,(M\,(2+\theta S)+(1+\theta S)\,(4-\theta BC+2\,\theta S))\right)\Big/((1+M+\theta S)\,(2+M-2\,\theta BC+2\,\theta S))+$$
$$\left(M\left(M^2+M\,(3+2\,\theta BC+\theta S)+2\left(3+4\,\theta S+\theta S^2\right)\right)\right)\Big/$$
$$((M+2\,\theta BC)\,(1+M+\theta S)\,(2+M+2\,\theta S))\Bigg)\Big/((1+\theta BC)\,(3+\theta S))$$

```
ilt2typesINCON2REV = InverseLaplaceTransform[Λ^-1 limSolGenINCON2REV, Λ, T] // Simplify
```

$$\left(\frac{2\, e^{-\frac{1}{2} T (2+M+2\,\theta S)} (1+\theta S)(-1-2\,\theta AB+\theta S)}{(1-\theta AB+\theta S)(2+M+2\,\theta S)} + \frac{M^2 (3+M+2\,\theta AB+\theta S)}{(M+2\,\theta AB)(1+M+\theta S)(2+M+2\,\theta S)} - \left(2\, e^{-\frac{1}{2} T (M+2\,\theta AB)} M\, \theta AB\, (3+\theta S)\right)\Big/ ((M+2\,\theta AB)(-1+\theta AB-\theta S)(2+M-2\,\theta AB+2\,\theta S)) + \frac{2\, e^{-T (1+M+\theta S)} (M+(2+\theta AB)(1+\theta S))}{(1+M+\theta S)(2+M-2\,\theta AB+2\,\theta S)}\right)\Big/ ((1+\theta AB)(3+\theta S))$$

To obtain the GF for topologically uninformative blocks we need to sum over all three possible topologies

```
limSolGenNOTOPbcREV = Limit[asymGFREV /. {ω[{a, b}] → z α, ω[{a, c}] → z, ω[{b, c}] → θ/2}, z → ∞] /.
   {ω[{c}] → ωsh - ω[{b}], ω[{a}] → ωa} // Simplify;
limSolGenNOTOPabREV = Limit[asymGFREV /. {ω[{a, c}] → z α, ω[{b, c}] → z, ω[{a, b}] → θ/2}, z → ∞] /.
   {ω[{a}] → ωsh - ω[{b}], ω[{c}] → ωc} // Simplify;
limSolGenNOTOPacREV = Limit[asymGFREV /. {ω[{a, b}] → z α, ω[{b, c}] → z, ω[{a, c}] → θ/2}, z → ∞] /.
   {ω[{a}] → ωsh - ω[{c}], ω[{b}] → ωb} // Simplify;
```

To GF for discrete splitting times are:

```
ilt2typesNOTOPbcREV = InverseLaplaceTransform[Λ^-1 limSolGenNOTOPbcREV, Λ, T] // Simplify;
ilt2typesNOTOPabREV = InverseLaplaceTransform[Λ^-1 limSolGenNOTOPabREV, Λ, T] // Simplify;
ilt2typesNOTOPacREV = InverseLaplaceTransform[Λ^-1 limSolGenNOTOPacREV, Λ, T] // Simplify;
```

Setting all ω to zero the GF has to sum to one:

```
asymGFREV /. {ω[_] -> 0} // Simplify
```

```
1
```

Topological probabilities sum to one as they should:

```
{topconREV = ilt2typesCON2REV /. {θS → 0,  θBC → 0},
  topinconREV = ilt2typesINCON2REV /. {θS → 0,  θAB → 0}} // Simplify
```

$$\left\{\frac{4\, e^{-\frac{1}{2}(2+M)\, T} - \frac{4\, e^{-(1+M)\, T}(2+M)}{1+M} + \frac{6+3\, M+M^2}{1+M}}{3\,(2+M)},\ \frac{1}{3}\left(\frac{2\, e^{-(1+M)\, T}}{1+M} - \frac{2\, e^{-\frac{1}{2}(2+M)\, T}}{2+M} + \frac{M\,(3+M)}{(1+M)\,(2+M)}\right)\right\}$$

```
topconREV + 2 topinconREV // FullSimplify
```

```
1
```

Wang & Hey reanalysis

Importing the data

This imports the Wang & Hey alignments (30,247 loci). The Dsim1/Dsim2/Dmel triplets have been filtered as described in W&H and polarized relative to Dyak. Divergent sites that are invariant in the ingroup are denoted as {1,1,1}, sites with more than two states (either due to backmutation or recombination) are denoted as {1,2,2}, {1,2,3} etc. Sites that are monomorphic in in and outgroup have been stripped, i.e. the order of mutation is retained, the sequence length information is lost. The file is still large (10 Mb):

```
WangHeyRaw = Partition[Import["/home/konrad/Downloads/ALLstripped2", "Table"], 3];
WangHeyRaw // Length
```

```
30 247
```

This turns alignment into lists of site types. The first locus is:

```
WangHeyRaw2 = sitetyp[WangHeyRaw]; WangHeyRaw2〚1〛
```

```
{{1, 1, 1}, {0, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 0, 0},
 {0, 0, 1}, {1, 0, 0}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 2, 2}, {1, 1, 1}, {1, 1, 1},
 {1, 1, 1}, {1, 1, 1}, {0, 1, 0}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 0, 0}, {1, 1, 1},
 {1, 1, 1}, {1, 1, 1}, {0, 1, 0}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1},
 {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 0, 0}, {1, 1, 1}, {1, 1, 1},
 {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1},
 {1, 1, 1}, {1, 1, 1}, {1, 0, 0}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 0, 0}}
```

E.g.the first locus has 55 variable sites in total, 10 of which are divergent between in and outgroup:

```
{WangHeyRaw2〚1〛 // Length, Count[WangHeyRaw2〚1〛, {1, 1, 1}]}
```

```
{55, 44}
```

Trimming

To keep the number of mutations per block manageable, account for mutational heterogeneity and to minimize the effect of intralocus recombination we will trim each locus to the same outgroup distance. Cutting after 16 divergent (between Dmel and Dyak) sites corresponds to roughly one third of the mean number of divergent sites in the full dataset. There are 2090 loci that fall below this cut-off, i.e. are not informative enough will be ignored:

```
WangHeyTrimRaw = DeleteCases[divcutter[16, #] & /@ WangHeyRaw2, {}];
```

We can the simply count the 6 different mutational types at each locus (in the following order {{1,0,1},{0,1,0},{1,1,0},{0,0,1},{0,1,1},{1,0,0},{1,1,1}}. Sites with multiple segregating states are ignored. Below the counts for the first locus, which only contains one internal mutations on the branch between Dmel and the two Dsim samples:

```
WangHeyTrimCounts = counttyp[#] & /@ WangHeyTrimRaw; WangHeyTrimCounts〚1〛
```

```
{0, 1, 0, 1, 1, 2, 13}
```

As expected by symmetry, the mean number of mutations on internal branches corresponding to the two different incongruent genealogies (1st and 3th value below) is the same:

```
Table[Mean[#〚i〛 & /@ WangHeyTrimCounts] // N, {i, 1, 7}]
```

```
{0.126327, 0.662215, 0.122811, 0.628689, 1.81937, 3.03601, 11.8461}
```

17% of loci have no topologically informative mutations:

```
Count[(Plus @@ {#〚1〛, #〚3〛, #〚5〛}) & /@ WangHeyTrimCounts, 0] / (WangHeyTrimCounts // Length) // N
```

```
0.174663
```

For the triplet analysis conflicting (in terms of the topology) shared derived mutations (i.e. on internal branches) in the same locus are not possible. However, in the W&H data this is the case for 14% of loci:

```
Mean[If[Count[#〚{1, 3, 5}〛, 0] < 2, 1, 0] & /@ WangHeyTrimCounts] // N
```

```
0.139823
```

First, we will remove blocks that have more than 2 topologically conflicting mutations (2.2%) . This filters out dubious alignments without biasing against the tails of the coalescence time distribution:

```
WangHeyTrimCounts2 =
  DeleteCases[(If[(Plus @@ Delete[Sort[#〚{1, 3, 5}〛], -1]) < 2, #, r]) & /@ WangHeyTrimCounts, r];
(Length[WangHeyTrimCounts] - Length[WangHeyTrimCounts2]) / Length[WangHeyTrimCounts] // N
```

```
0.0226231
```

Second, we will assume that single incongruent site are backmutations and remove those from each alignment. For 5% of loci, the incongruent, i.e. less frequent topological site cannot be determined (because there are exactly two conflicting shared derived mutations), these loci will be removed:

```
Count[(Delete[Sort[#⟦{1, 3, 5}⟧], 1]) & /@ WangHeyTrimCounts2, {1, 1}] /
  Length[WangHeyTrimCounts2] // N
```

```
0.050109
```

```
{Count[#⟦{1, 3, 5}⟧ & /@ WangHeyTrimCounts2, {1, 1, 0}],
 Count[#⟦{1, 3, 5}⟧ & /@ WangHeyTrimCounts2, {1, 0, 1}],
 Count[#⟦{1, 3, 5}⟧ & /@ WangHeyTrimCounts2, {0, 1, 1}]}
```

```
{112, 659, 608}
```

```
WangHeyTrimCounts3 = DeleteCases[(incontrim3[#]) & /@ WangHeyTrimCounts2, r];
```

The mean number of mutations is reduced by more then half due to these trimming steps:

```
{Table[Mean[#⟦i⟧ & /@ WangHeyTrimCounts2] // N, {i, 1, 7}],
 Table[Mean[#⟦i⟧ & /@ WangHeyTrimCounts3] // N, {i, 1, 7}]}
```

```
{{0.101272, 0.64335, 0.096439, 0.609012, 1.81192, 3.04895, 11.8968},
 {0.0436479, 0.622279, 0.0403963, 0.588271, 1.85261, 3.0674, 11.9347}}
```

The number of loci removed in the various trimming steps is comparatively small. The only drastic reduction occurs when trimming to a fixed outgroup distance.

```
Length[WangHeyTrimCounts] - Length[WangHeyTrimCounts3]
```

```
2016
```

```
{Length[WangHeyRaw], Length[WangHeyTrimCounts],
 Length[WangHeyTrimCounts2], Length[WangHeyTrimCounts3]}
```

```
{30 247, 28 157, 27 520, 26 141}
```

■ Tests on pairwise data

It is quickest to run pairwise analyses (one Dmel, one Dsim individual) to compare the effect of the various trimming steps on parameter estimation and check against the W&H estimates.

▫ *Full dataset*

This throws out one of the Dsim individuals and condenses the data into counts of pairwise differences within the ingroup (S_in) and between ingroup and outgroup (S_out). Sites with more than two states (backmutations) are counted both in S_in and S_out, so the only difference to the W&H analysis is that we are fitting simpler IM models (with only one migration rate) and are assuming infinite sites mutations.

```
WangHeyPairs = topair[#] & /@ WangHeyRaw2; WangHeyPairs⟦1⟧
```

```
{10, 51}
```

We need to tabulate LogL of M and T for all observed values of S_in and S_out. There are 79*260= potential combinations.

```
{Table[Max[#⟦i⟧ & /@ WangHeyPairs], {i, 1, 2}],
 Table[Min[#⟦i⟧ & /@ WangHeyPairs], {i, 1, 2}], Table[Mean[#⟦i⟧ & /@ WangHeyPairs] // N, {i, 1, 2}]}
```

```
{{79, 260}, {0, 0}, {18.0691, 46.5621}}
```

Scaling locus specific mutation rates based on the number of observed S_out values and tabulating all LogL exactly would take very long,; a much faster alternative is to bin contigs according to their outgroup divergence, 10 bins should be enough:

```
tabu = Table[Select[ WangHeyPairs, #⟦2⟧ > (260 / 10) * i && #⟦2⟧ < (260 / 10) * (i + 1) &], {i, 0, 9}];
bincounts = Table[Table[Count[#⟦1⟧ & /@ tabu⟦i⟧, k], {k, 0, 79}], {i, 1, 10}];
```

The mutation rate scalars (relative to the mean divergence across all blocks) for the bins are:

```
meanmut = Mean[WangHeyPairs] // N;
meanbin = Table[Mean[#⟦2⟧ & /@ tabu⟦i⟧] // N, {i, 1, 10}] / meanmut⟦2⟧
```

```
{0.385613, 0.816086, 1.35804, 1.90756, 2.46236, 3.0201, 3.60045, 4.12813, 4.69419, 5.16514}
```

The joint MLE for τ and θ under a simple split model without migration are:

```
splitMLEFull = FindMaximum[
  Plus @@ Table[Total[Table[Log[Psplit[τ, θ, {meanbin⟦i⟧, k}]], {k, 0, 79}] * bincounts⟦i⟧],
    {i, 1, 10}], {τ, 0.5, 0.1, 4}, {θ, 8, 4, 16}]

{-94119.3, {τ → 2.18337, θ → 5.82369}}
```

The joint MLE for M, τ and θ for IM model with symmetric and asymmetric migration are:

```
imMLEFull = FindMaximum[Plus @@
   Table[Total[Table[Log[WilkHeSim2s[M, τ, θ, {meanbin⟦i⟧, k}]], {k, 0, 79}] * bincounts⟦i⟧],
    {i, 1, 10}], {M, 0.05, 0, 0.5}, {τ, 0.5, 0.1, 3}, {θ, 8, 4, 16}]

{-93467.4, {M → 0.0256439, τ → 2.69317, θ → 5.14413}}

imMLEFullasym = FindMaximum[
  Plus @@ Table[Total[Table[Log[asym2s[M, τ, θ, {meanbin⟦i⟧, k}]], {k, 0, 79}] * bincounts⟦i⟧],
    {i, 1, 10}], {M, 0.05, 0, 0.5}, {τ, 0.5, 0.1, 3}, {θ, 8, 4, 16}]

{-93466.3, {M → 0.0510174, τ → 2.69555, θ → 5.14185}}
```

Note that the MLE for M under the symmetric model is half that inferred for the asymmetric migration model as expected.

▫ *Trimmed to fixed divergence*

Repeating the above for the data (without trimming out backmutations and topologically conflicting mutations):

```
WangHeyTrimPairs2 = topair[#] & /@ WangHeyTrimRaw;
{Mean[WangHeyTrimPairs2] // N, Max[(#⟦1⟧ & /@ WangHeyTrimPairs2)]}

{{6.45857, 15.9826}, 31}

tabPairs = Table[Count[(#⟦1⟧ & /@ WangHeyTrimPairs2), i], {i, 0, 31}];
splitMLE = FindMaximum[
  Total[Table[Log[Psplit[τ, θ, {1, i}]], {i, 0, 31}] * tabPairs], {{τ, 0.2, 0, 4}, {θ, 2, 1, 4}}]

{-70303.4, {τ → 3.07509, θ → 1.58489}}
```

There is still a clear signal of migration (M is slightly lower than in the analysis on the full data):

```
imMLEasym = FindMaximum[Total[Table[Log[asym2s[M, τ, θ, {1, i}]], {i, 0, 31}] * tabPairs],
  {M, 0.02, 0, 0.5}, {τ, 3.5, 0.5, 4}, {θ, 2, 0.9, 4}]

{-70180., {M → 0.041833, τ → 3.84235, θ → 1.37638}}
```

▫ *Trimmed to fixed divergence, no backmuts and incongruent sites*

What effect does ignoring detectable backmutations and conflicting shared derived mutations have?

```
WangHeyTrimPairs3 = (Plus @@ Drop[Drop[#, 1], -1]) & /@ WangHeyTrimCounts3;
tabPairs3 = Table[Count[WangHeyTrimPairs3, i], {i, 0, Max[WangHeyTrimPairs3]}];

imMLE3asym =
 FindMaximum[Total[Table[Log[asym2s[M, τ, θ, {1, i}]], {i, 0, Max[WangHeyTrimPairs3]}] * tabPairs3],
  {M, 0.1, 0.001, 0.6}, {τ, 3, 0.5, 6}, {θ, 2, 0.9, 3}]

{-65717.1, {M → 0.0933362, τ → 3.33526, θ → 1.50898}}
```

■ Triplet analysis

Given the symmetry in the model there are only 3 types of loci, congruent, incongruent and those without parsimony informative sites. Within each class there are 3 types of mutations, those on the shorter external branches, those on the internal branch and those on longer external branch (the counts are listed in this order). The function sitecount sorts loci according to topology. The mutational information at each locus is summarized by counting the number of mutations on each branch. The first locus with a congruent topology has 2 mutations on the shorter external branches, one on the internal branch and one on the longer external branch.

```
WangHeyTrimCounts3 // Length
```

```
26141
```

```
WHcount = sitecount3s[WangHeyTrimCounts3]; WHcount⟦1, 1⟧
```

```
{2, 2, 1}
```

To make the GF calculation feasible we need to exclude 6 extreme loci with very large numbers of mutations (>16) on anyone branch. This should have very little effect on parameter estimates but avoids catastrophic rounding error. We can then summarize the data as counts of distinct mutational configurations in each topology class:

```
WHCount2 = {Select[WHcount⟦1⟧, (Max[#]) < 17 &], Select[WHcount⟦2⟧, (Max[#]) < 15 &],
   Select[WHcount⟦3⟧, (Max[#]) < 14 &]}; max2 = maxcount3s[WHCount2]
```

```
{{16, 12, 16}, {12, 13, 8}, {8, 13, 11}}
```

Note that the most diverse locus still has 26 mutations.

```
Table[Max[Total[#] & /@ (WHCount2⟦i⟧)], {i, 1, 3}]
```

```
{26, 19, 18}
```

```
resWH2 = Table[Table[Count[WHCount2[[r]], {i, j, k}],
     {i, 0, max2[[r, 1]]}, {j, 0, max2[[r, 2]]}, {k, 0, max2[[r, 3]]}], {r, 1, 3}];
```

```
resWH2 // Flatten // Total
```

```
26135
```

How to best tabulate the probabilities of the observed configurations? The simplest approach is to tabulate the probabilities for all possible configurations (given the maximum number of mutations observed on each branch).

The function tripletL computes LogL under the IM model with asymmetric migration. For a single point in parameter space this takes about 1.5 seconds:

```
Timing[tripletL[0.16, 3.3, 1.5, resWH2, max2]]
```

```
{1.45609, -149746.}
```

FindMaximum uses derivatives and finds the MLE estimate in a few minutes:

```
Timing[triplMax =
   FindMaximum[tripletL[M, τ, θ, resWH2, max2], {M, 0.1, 0.001, 0.6}, {τ, 2, 1, 6}, {θ, 1, 1, 3}]]
```

```
{439.859, {-149556., {M → 0.173665, τ → 3.34091, θ → 1.39874}}}
```

▫ *Comparison between sampling schemes and with Wang and Hey*

How do the above MLEs compare to the estimates of W&H. Given that there are various differences in the scaling of parameters (W&H scale both divergence and migration relative to the mutation rate), we need to convert these into absolute values. W&H assume that Dmel and Dyak diverged 10 MYA with 10 generations per year.The μ per locus and generation for the full data and the fixed divergence are:

```
{{imMLEFull[[2, 3, 2]], imMLEFull[[2, 2, 2]], imMLEFull[[2, 1, 2]]},
  {imMLEFullasym[[2, 3, 2]], imMLEFullasym[[2, 2, 2]], imMLEFullasym[[2, 1, 2]]},
  {imMLEasym[[2, 3, 2]], imMLEasym[[2, 2, 2]], imMLEasym[[2, 1, 2]]},
  {imMLE3asym[[2, 3, 2]], imMLE3asym[[2, 2, 2]], imMLE3asym[[2, 1, 2]]} ,
  {triplMax[[2, 3, 2]], triplMax[[2, 2, 2]], triplMax[[2, 1, 2]]}} // TableForm
```

```
5.14413    2.69317    0.0256439
5.14185    2.69555    0.0510174
1.37638    3.84235    0.041833
1.50898    3.33526    0.0933362
1.39874    3.34091    0.173665
```

```
mufull = 0.1 * (meanmut⟦2⟧ / 2) / 10000000 // N; mu1 = 0.1 * (16 / 2) / 10000000 // N;
```

Converting into absolute values is straightforward for T and N_e. Below the MLE estimates for model parameters for the full data(2nd column), the length trimmed data (3rd column), length trimmed data without backmutations and incongruent sites (4th column) and the same data:

```
Nefs = imMLEFull⟦2, 3, 2⟧ / (4 * mufull);
Tfs = 0.1 * imMLEFull⟦2, 2, 2⟧ * 2 * Nefs;
mfs = imMLEFull⟦2, 1, 2⟧ * (2.44 * 10 ^ 6) / Nefs;

Nef = imMLEFullasym⟦2, 3, 2⟧ / (4 * mufull);
Tf = 0.1 * imMLEFullasym⟦2, 2, 2⟧ * 2 * Nef;
mf = imMLEFullasym⟦2, 1, 2⟧ * (2.44 * 10 ^ 6) / (2 Nef);

Ne1 = imMLEasym⟦2, 3, 2⟧ / (4 * mu1);
T1 = 0.1 * imMLEasym⟦2, 2, 2⟧ * 2 Ne1;
m1 = imMLEasym⟦2, 1, 2⟧ * (2.44 * (0.00513 / 0.0055) * 10 ^ 6) / (2 Ne1);

Ne3 = imMLE3asym[[2, 3, 2]] / (4 * mu1);
T3 = 0.1 * imMLE3asym[[2, 2, 2]] * 2 Ne3;
m3 = imMLE3asym[[2, 1, 2]] (2.44 * (0.00513 / 0.0055) * 10 ^ 6) / (2 Ne3);

Netr = triplMax⟦2, 3, 2⟧ / (4 * mu1);
Ttr = 0.1 * triplMax⟦2, 2, 2⟧ * 2 Netr;
mtr = triplMax⟦2, 1, 2⟧ (2.44 * (0.00513 / 0.0055) * 10 ^ 6) / (2 Netr);
{{Nefs, Nef, Ne1, Ne3, Netr }, {Tfs, Tf, T1, T3, Ttr}, {mfs, mf, m1, m3, mtr}} // TableForm
```

```
5.52395 × 10^6    5.5215 × 10^6    4.30119 × 10^6    4.71556 × 10^6    4.37105 × 10^6
2.97538 × 10^6    2.9767 × 10^6    3.30533 × 10^6    3.14553 × 10^6    2.92065 × 10^6
0.0113272         0.0112725        0.0110674         0.0225232         0.0452107
```

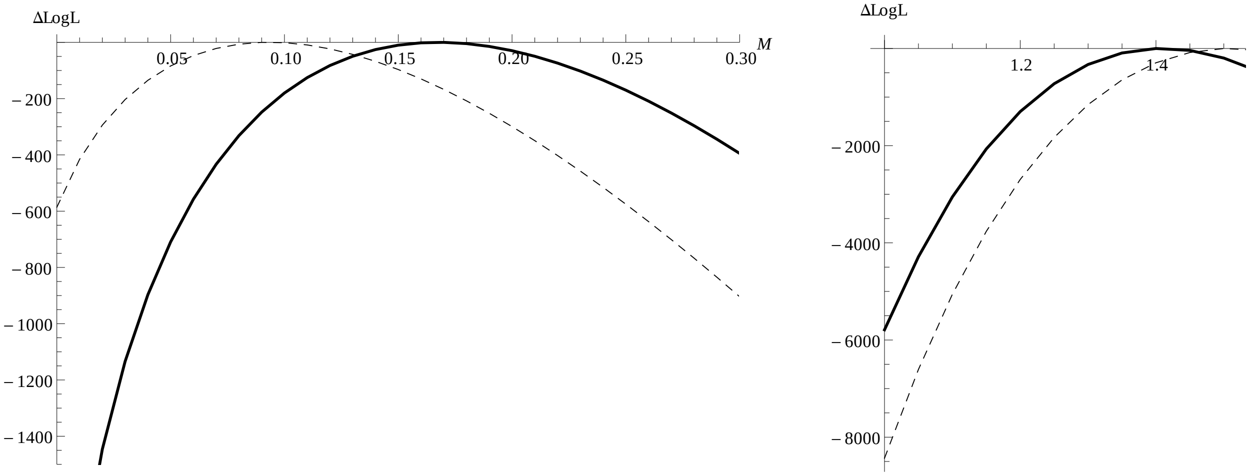
The effective population size and divergence time estimates in the pairwise analysis on the full data agree very well with those of W&H. The effective pop. size is slightly larger than the ancestral N_e estimated by W&H but smaller than their estimate of the Dsim effective population size. Given that our simpler model only contains one N_e parameter, one would expect the MLE for this parameter to be inbetween that of the ancestral population and Dsim.

Note that W&H scale migration as $M=2N_{dmel}\ m$, we are scaling M=4Nanc m. If we taking the larger effective pop. size of the ancestral population compared to Dmel (2.44 Mio) into account, M matches the W&H estimate (0.013) quite well. However, ignoring backmutations and incongruent mutations within blocks results in a marked increase in M and a decrease in N_e , which makes sense given that we are removing polymorphic sites. Strikingly, in the triplet analysis, the estimate of M is further increased compared to the pairwise analysis on the same dataset.

■ Comparing pairwise and triplet results

To visualize the difference between pairwise and triplet estimates, we evaluate a profile through the maximum of the likelihood surface for each parameter (fixing the other two parameters at their MLE):

The ΔLogL (relative to the maximum) for M, T and θ for pairwise (dashed line) and triplet (solid line) reveal not only that the MLE differ between the two sampling schemes (M is higher, θ lower in the triplet analysis), but also that the curvature is the same, i.e. there is no improvement in power by adding a 3rd sample which is unexpected.



What explains the difference between the pairwise and the triplet estimates (in terms of bias and power)? The triplet estimates (in particular M) should be sensitive to any model violation (both the model of sequence evolution and history) that affects the inferred frequencies of incongruent topologies. For relatively old divergence (as here) most incongruent genealogies are expected to be due to migration rather than incomplete lineage sorting. We can use the GF to find the expected frequencies of the three topology classes (congruent, incongruent and uninformative, see Table below) given the MLE for the two sampling schemes and compare these against the observed frequencies. The expected frequency of observable blocks with a congruent topologies is given by the frequency of the congruent genealogies (minus the proportion of those in that have no shared derived mutations).

```
0.821639    0.0211294    0.157232
0.772016    0.0286075    0.199377
0.750067    0.0617563    0.188177
```

There is an excess of incongruent topologies in the data (6.1%), which cannot be explained by the inferred histories. However, the observed frequencies (last row above) match the expectations from triplet MLEs (middle row) much better than those corresponding to the pairwise analysis (1st row).

Given the frequency of sites with more than 2 segregating sites, backmutations in the outgroup branch (which lead to mispolarized sites) are the most likely explanation. To check this we can look at the average number of mutations on each branch in the 3 topology classes. While congruent loci have on average fewer mutations on the two shorter external branches (i.e. those leading to the common ancestor of the two Dsim individuals) (1st row, 1st column) than on the longer external branch (2nd column); this is not the case at all for incongruent loci (2nd row). Thus most loci inferred to have an incongruent topology are due to mispolarized mutations. Given the magnitude of the excess of incongruent loci, it is actually surprising how well the triplet scheme still works!

```
Table[WHCount2⟦i⟧ // Mean // N, {i, 1, 3}] // TableForm
```

```
1.15411     3.19507     2.46722
3.39715     1.15551     1.35812
0.562627    0.563847    2.76759
```

■ Comparison with simulated data

This imports 26141 loci simulated for triplet sampled {{b,c},a} simulated under the IM model with asymmetric migration using Hudson's ms. The values used for simulation were those inferred in the pairwise analysis (T= 3.33, M=0.0933, θ=1.5). The key question is how much statistical power can be gained from analyzing triplet samples compared to pairs ?

```
sim = ReadList["/home/konrad/Downloads/WangHeyTest3rd"]⟦1⟧; Mean[sim] // N
```

```
{0.0178647, 0.719559, 0.020619, 0.727095, 2.33863, 3.03925}
```

Around 84% of the loci are topologically informative:

```
Total[If[#⟦1⟧ > 0 || #⟦3⟧ > 0 || #⟦5⟧ > 0, 1, 0] & /@ sim] / Length[sim] // N
```

```
0.841322
```

```
res = sitecount3s[sim]; max = maxcount3s[res]
```

```
{{14, 15, 16}, {16, 8, 8}, {9, 10, 11}}
```

This summarizes loci in each topological class as counts of distinct mutational configurations:

```
res2 = configcount3s[sim];
```

In contrast to the real data, both the pairwise and triplet results closely match the true values used for simulations:

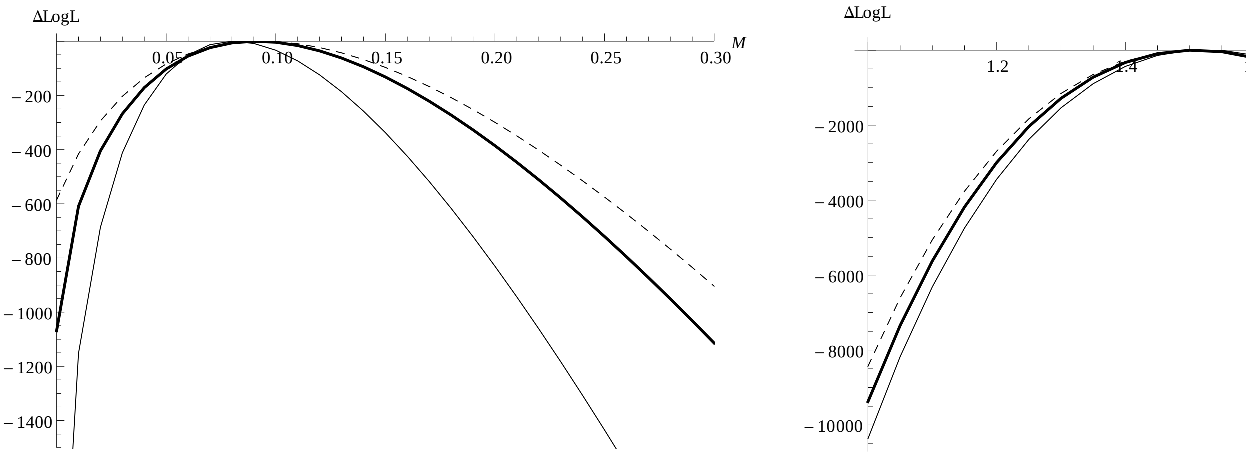
```
simpair = (Flatten[Delete[#, {{1}, {4}}]] // Total) & /@ sim;
simcount = Table[Count[simpair, i], {i, 0, Max[simpair]}];
pairsim = FindMaximum[Total[Table[Log[asym2s[M, τ, θ, {1, i}]], {i, 0, Max[simpair]}] * simcount],
  {M, 0.1, 0.001, 2}, {τ, 4, 1, 12}, {θ, 1, 0.1, 2}]
```

```
{-65619.2, {M → 0.0984103, τ → 3.23947, θ → 1.53159}}
```

```
Timing[triplsim =
  FindMaximum[tripletL[M, τ, θ, res2, max], {M, 0.1, 0.01, 0.5}, {τ, 3.3, 1, 6}, {θ, 1.5, 1, 4}]]
```

```
{376.844, {-151483., {M → 0.0922354, τ → 3.28638, θ → 1.51179}}}
```

This plots the difference in LogL from its maximum (ΔLogL) against T (left) and θ (right) for triplet (solid, thick lines) and pairwise (dashed lines) samples. As expected and in contrast to the inference on the real data, the triplet estimates are narrower. If one uses the reverse triplet sampling scheme i.e. sampling two individuals from the receiving population (see analysis of data simulated for this case with the same parameter values below), the power to infer M increases substantially (thin solid lines):



This imports data simulated under the reverse sampling scheme:

```
simREV = ReadList["/home/konrad/Downloads/WangHeyTestREV"][[1]]; Mean[simREV] // N
```

{{1, 0, 1}, {0, 1, 0}, {1, 1, 0}, {0, 0, 1}, {0, 1, 1}, {1, 0, 0}}

```
{0.0881757, 0.826556, 0.0827053, 0.819555, 2.2545, 3.01067}
```

```
resREV = sitecount3s[simREV]; max = maxcount3s[resREV]
```

```
{{14, 15, 15}, {17, 10, 11}, {10, 10, 11}}
```

```
resREV = configcount3s[simREV];

Timing[tripletLREV[0.0933, 3.3, 1.5, resREV, max]]

{1.48409, -158940.}
```

Finding the Maximum takes 15 mins....

```
Timing[triplsimREV = FindMaximum[tripletLREV[M, τ, θ, resREV, max],
   {M, 0.1, 0.01, 0.5}, {τ, 3.3, 1, 6}, {θ, 1.5, 1, 4}]]

{830.436, {-158920., {M → 0.0868254, τ → 3.26594, θ → 1.51019}}}
```

```
Neps = pairsim⟦2, 3, 2⟧ / (4 * mu1);
Tps = 0.1 * pairsim⟦2, 2, 2⟧ * 2 * Neps;
mps = pairsim⟦2, 1, 2⟧ * (2.44 * 10 ^ 6) / (2 Neps);

Netrs = triplsim⟦2, 3, 2⟧ / (4 * mu1);
Ttrs = 0.1 * triplsim⟦2, 2, 2⟧ * 2 * Netrs;
mtrs = triplsim⟦2, 1, 2⟧ * (2.44 * 10 ^ 6) / (2 Netrs);
```

```
{triplsim[[2, 1, 2]], pairsim⟦2, 1, 2⟧}

{0.0922354, 0.0984103}

{{Neps, Netrs }, {Tps, Ttrs}, {mps, mtrs}} // TableForm

4.78623×10^6     4.72436×10^6
3.10097×10^6     3.10521×10^6
0.0250846        0.0238185
```

3. Numbers of configurations

The feasibility of finding a solution for the GF depends on the number of configurations that need to be tracked. In a two-deme migration model, the number of configurations that are possible is determined by the number of ways that j lineages present between successive coalescent events can be distributed across the two populations, and the number of ways that the ancestry of n sampled individuals can be distributed amongst the j lineages present in each successive coalescence event. Specifically, the total number of configurations is:

$$\sum_{j=2}^{n} 2\,(S_{j,2}+1)\,S_{n,j} \qquad (1)$$

where $S_{n,j}$ is the Stirling number of the second kind, which gives the number of ways that n lineages can be distributed over j non-empty sets. The sum is over all the intervals during which there were j extant lineages. This number grows dramatically with the number of lineages. For example, there are 92 and 2428 configurations for $n = 4$ and 6 respectively. In the IM model there are an additional $\sum_{j=2}^{n} S_{n,j}$ configurations possible in the ancestral population.

However, if we can find algebraic expressions for the GF with j genes, we do not need to track all these configurations: for example, if we know $\psi[a, b, c \setminus \emptyset]$, we can immediately find $\psi[a\,b, c, d \setminus \emptyset]$, for example. Therefore, the number of types of configuration that we need to calculate is only:

$$\sum_{j=2}^{n} 2\,(S_{j,2}+1) = 2^{n+1} - 4 \qquad (2)$$

For example, this is 28 and 124 for $n = 4$ and 6, respectively.

Although the numbers of configurations with (say) 6 genes would be manageable for numerical calculations, extracting probabilities of mutational configurations requires that we differentiate an algebraic expression, which is given by inverting a large matrix. However, as discussed above the GF can be found directly if it is written as an expansion in M or R, each term corresponding to histories with 0, 1, ...

migration or recombination events. The question is now, how many different histories do we need to track, if we allow k mutation or recombination events?Consider migration between two demes. A migration event can occur in j ways during the interval when there are j lineages, and so a single event can occur in $n + (n - 1) + \ldots + 2 = (n + 2)(n - 1)/2$ ways. Multiple events occur independently, and so there are $((n + 2)(n - 1)/2)^k$ ways that k migration events can occur in the history of n genes. For example, with 4 genes there are 9, 81, 729 ways that 1, 2, 3 migration events can occur, and with 6 genes there are 20, 400, 8000 terms, respectively. With this method, we need to track many more configurations, but each is given by a much more direct calculation.

If we observe a very large number of loci, then we wish to tabulate the probability of every possible configuration of mutations. With n genes, there are $2^n - 2$ branches, and so we have $(2^n - 2)^k$ ways to throw down k mutations onto the branches. For example, even with 3 genes there are 6 possible branches, and $6^{10} \sim 6 * 10^6$ ways to distribute 10 mutations over the branches. However, the number of possibilities that we need to tabulate is much smaller than this, because the probability is determined by a much smaller number of sufficient statistics. With three genes, if we observe no mutations on the internal branches, then the probability depends only on the numbers of singletons, $\{k_a, k_b, k_c\}$, whilst if we see (say) at least one mutation ancestral to $\{a, b\}$, then we know the topology: then, the probability is determined by $\{k_a + k_b, k_{a b}, k_c\}$. In both cases, there are 14 distinct ways to divide 10 mutations over 3 classes of mutation. With more genes, more classes must be tabulated, but their number does not increase catastrophically. For example, with 6 genes and no internal mutations, there are 35 ways to distribute 10 mutations across 6 singleton classes. At the other extreme, if we know the topology, then the probability is determined by 5 independent coalescence times, and so we expect that tabulating the probability of getting i_1, …i$_5$ mutations in each of the five time intervals will allow us to calculate the chance of seeing a particular set of mutations $\underline{k}$. All except the singleton class must contain mutations, and so there are roughly $\sum_{k=1}^{10} \sum_{j=1}^{k} \sum_{i=1}^{j} \sum_{l=0}^{i} 1 = 935$ configurations.

Definitions

■ **Automating the recursions**

□ ***makeEqns***

□ ***makeAllEqns***

□ ***TotalRate***

□ ***Mergers***

□ ***reduceEqns***

□ ***tidyNotation***

□ ***numberOfDemes***

□ ***numberOfGenes***

□ ***GetVars***

□ ***configs***

□ ***selectEqns***

■ **Data analysis**

□ ***sitecount3s***

□ ***configcount3s***

□ ***maxcount3s***

□ ***pr3s***

□ ***tripletL***

□ ***WilkHeSim2s***

□ ***asym2s***

- ***divcutter***
- ***sitetyp***
- ***topair***
- ***counttyp***
- ***incontrim3***
- ***Psplit***
- ***probSasym***