One of the key properties of Kingman’s coalescent is that each pair of lineages is equally likely to coalesce whenever a coalescent event occurs. This condition is violated in structured populations, since lineages will on average be more closely related to other lineages belonging to the same subpopulation. Consequently, Kingman’s coalescent does not correctly describe the distribution of the genealogy of a random sample of chromosomes from a structured population. Instead, we must work with a generalization called the structured coalescent.
Example: Two Demes

Suppose that a population of diploid organisms is subdivided into two demes, where:

- The population sizes are $N_1$ and $N_2$, respectively.
- Generations are non-overlapping.
- Reproduction in each deme follows the Wright-Fisher model, modified to allow for migration.
- Specifically, a chromosome in deme $i$ will either be descended from a randomly sampled chromosome from the same deme with probability $m_{ii}$ or from a randomly sampled chromosomes from the other deme with probability $m_{ji}$. 
Backward vs. Forward Migration Rates

The quantities $m_{ii}$ and $m_{ij}$ introduced on the previous slide are called **backward migration rates** since they describe the rate at which lineages in deme $i$ either remain in deme $i$ or move to deme $j$ **when we are looking backwards in time**.

In contrast, in ecology, we usually think in terms of **forward migration rates**, which describe the rate at which individuals born in one location migrate to a new location. Under our modified Wright-Fisher model, these are related as follows. If $q_{ij}$ is the probability that an individual born in deme $i$ migrates to deme $j$, then

$$m_{ij} = \frac{N_j}{N_i} q_{ji}.$$  

**Caveat:** Most coalescent-based analyses of migration report backward migration rates.
Scaling Assumptions

It is customary to make the following assumptions.

1. We assume that the population sizes $N_1$ and $N_2$ are large enough (say $> 100$) that multiple mergers and other complex coalescent events can be neglected (i.e., the genealogy is almost certain to be a binary tree).

2. We also assume that the backwards migration rates are of order $(N_1 + N_2)^{-1}$. This ensures that it is very unlikely that two or more events will happen in the same generation.

Remark: If the second condition is not satisfied, then migration events will happen much more rapidly than coalescent events so that the population is effectively panmictic on the coalescent time scale. In this case, Kingman's coalescent applies with a suitably defined $N_e$. 
Suppose that we have sampled $n_1$ chromosomes from deme 1 and $n_2$ chromosomes from deme 2. Looking backwards in time, the genealogy of this sample can be influenced by the following events:

- At rate $\left(\frac{n_1}{2}\right) \frac{1}{2N_1}$, a randomly chosen pair of lineages in deme 1 coalesces, reducing $n_1$ to $n_1 - 1$.
- At rate $\left(\frac{n_2}{2}\right) \frac{1}{2N_2}$, a randomly chosen pair of lineages in deme 2 coalesces, reducing $n_2$ to $n_2 - 1$.
- At rate $n_1 m_{12}$, a randomly chosen lineage moves from deme 1 to deme 2, reducing $n_1$ to $n_1 - 1$ and increasing $n_2$ to $n_2 + 1$.
- At rate $n_2 m_{21}$, a randomly chosen lineage moves from deme 2 to deme 1, reducing $n_2$ to $n_2 - 1$ and increasing $n_1$ to $n_1 + 1$.

This process continues until there is only one lineage remaining, at which point the entire genealogy of the sample is determined, including information about the locations of ancestral lineages. This is known as the **structured coalescent**.
Structured Coalescents

We can also define structured coalescents when there are more than two demes. Suppose that there are $D$ demes of sizes $N_1, \cdots, N_D$ and that the backward migration rate from deme $i$ to deme $j \neq i$ is $m_{ij}$. Given a random sample of $n_1$ chromosomes from deme 1, $n_2$ chromosomes from deme 2, etc., the following events can occur:

- **Coalescent events:** At rate $\binom{n_i}{2} \frac{1}{2N_i}$, a randomly chosen pair of lineages in deme $i$ coalesces, reducing $n_i$ to $n_i - 1$.

- **Migration events:** At rate $n_i m_{ij}$, a randomly chosen lineage moves from deme $i$ to deme $j$, reducing $n_i$ to $n_i - 1$ and increasing $n_j$ to $n_j + 1$. 
Pairwise Coalescent Times

Suppose that the population contains $D$ demes, each containing $N$ diploid individuals, and that the migration rate between any pair of demes is $m$. This is \textit{Wright’s finite island model}. Let $T_w$ and $T_b$ be the pairwise coalescent times (in generations) when two chromosomes are either sampled from the same deme or from two different demes. Then

\[
\mathbb{E}[T_w] = 2ND
\]
\[
\mathbb{E}[T_b] = 2ND + \frac{(D - 1)}{2m}
\]

- Notice that $\mathbb{E}[T_w]$ does not depend on $m$, but does depend on $D$, i.e., even if we sample chromosomes from the same deme, their expected coalescent time is affected by population structure.
- The pairwise coalescent time of two chromosomes sampled from different demes does depend on $m$ and is greater than that of two chromosomes sampled from the same deme.
Inference using the Structured Coalescent

Several software packages are available that use the structured coalescent to analyze sequence data sampled from subdivided populations.

- **MIGRATE** is a program by Peter Beerli (Beerli 2009) that can be used to estimate backward migration rates and effective population sizes from sequence data sampled from a subdivided population.
- It is based on the modified Wright-Fisher model described above.
- It can carry out both Bayesian and maximum likelihood inference (one at a time) using Monte Carlo algorithms.
- Substitution models: infinite alleles model; F84 model for DNA sequences; infinite sites model for SNP data; SMM model for microsatellites
- Inference can be done using multiple unlinked loci, but each locus is assumed to be non-recombining.
Bayesian Phylogeography using BEAST

BEAST is able to analyze data from subdivided populations using an approximation to the structured coalescent known as the discrete phylogeographical model (Lemey et al. 2009).

- In this approach, location is treated as a neutrally evolving character that takes values in a discrete set (the set of locations).
- Pairs of lineages having the same location can coalesce and individual lineages can move between locations.
- A variable selection procedure (stochastic search variable selection; SSVS) is used to sample sparse “migration rate matrices”, i.e., this method tends to minimize the number of “migration rates” that are positive.
- It is also possible to run analyses that use a GLM to identify relationships between the “migration rates” and other variables of interest such as temperature, transport networks, etc. (Lemey et al. 2014).
Regional migration rates of *P. vivax* inferred from mtDNA

- BEAST estimates both the inclusion probability and the magnitude of each rate.
- The inclusion probability is the posterior probability that the rate is non-zero.
- Inclusion probabilities can be converted to Bayes factors; customarily, a rate is reported only if $Bf > 3$.
- The magnitude is expressed as migrations per lineage per unit time.

<table>
<thead>
<tr>
<th>Region 1</th>
<th>Region 2</th>
<th>$P_{inc}$</th>
<th>BF</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>C America</td>
<td>SE Asia</td>
<td>0.519</td>
<td>3.389</td>
<td>0.695</td>
</tr>
<tr>
<td>C America</td>
<td>S Asia</td>
<td>0.512</td>
<td>3.296</td>
<td>0.843</td>
</tr>
<tr>
<td>S America</td>
<td>SE Asia</td>
<td>0.956</td>
<td>68.250</td>
<td>0.322</td>
</tr>
<tr>
<td>E Asia</td>
<td>SE Asia</td>
<td>1.000</td>
<td>Inf</td>
<td>0.787</td>
</tr>
<tr>
<td>SE Asia</td>
<td>S Asia</td>
<td>0.997</td>
<td>104.923</td>
<td>0.760</td>
</tr>
<tr>
<td>SE Asia</td>
<td>Melanesia</td>
<td>1.000</td>
<td>Inf</td>
<td>1.319</td>
</tr>
<tr>
<td>S Asia</td>
<td>Madagascar</td>
<td>1.000</td>
<td>Inf</td>
<td>2.884</td>
</tr>
<tr>
<td>S Asia</td>
<td>Africa</td>
<td>0.991</td>
<td>345.880</td>
<td>1.779</td>
</tr>
<tr>
<td>Melanesia</td>
<td>Africa</td>
<td>0.815</td>
<td>13.838</td>
<td>0.785</td>
</tr>
<tr>
<td>Melanesia</td>
<td>Middle East</td>
<td>0.990</td>
<td>310.978</td>
<td>0.768</td>
</tr>
</tbody>
</table>

Taylor et al. (2013)
Inference of Ancestral Locations in BEAST: Rabies in African Wild Dogs

Predictors of global H3N2 diffusion

Isolation-with-Migration Models

The structured coalescent can also be extended to models that allow the number of populations and their migration rates to change over time. These are known as isolation-with-migration models and are useful for analyzing biogeographical processes and incipient speciation.

- IM is a program by Hey & Nielsen (2007) that can be used to do inference for IM models using DNA data.
- Hey (2010) extended this to allow for arbitrary numbers of populations and phylogenies (IMa2).
- Inference is carried out within a Bayesian framework using MCMC.

Source: Hey (2010)
Assumptions

- The history of the sampled populations can be represented by a rooted bifurcating tree.
- No gene flow has occurred between unsampled populations and those contained within the tree.
- Each subpopulation is constant in size and follows the Wright-Fisher model.
- Pairwise migration rates are constant, but not necessarily symmetrical.
- No bottlenecks occurred when ancestral populations split.
- The individual loci are unlinked and no recombination occurs within loci.
- Each locus evolves neutrally: no selection at the locus or at other linked loci.
- Substitution models: HKY and infinite sites for sequence data; SMM for microsatellite data.
IMa2 Input and Estimated Parameters

Suppose that we have sequence data sampled from a population subdivided into $k$ subpopulations. To analyze this using IMa2, we need a bifurcating tree which describes a history of population splits terminating in a single ancestral population. The following unknown parameters will be estimated by the program.

- $k$ population mutation rates $\theta_1, \cdots, \theta_k$ for the extant populations and $k - 1$ population mutation rates $\theta_{k+1}, \cdots, \theta_{2k-1}$ for the ancestral populations. These are all scaled by the average mutation rate $\mu$.
- $2(k - 1)^2$ backward migration rates $m_{ij} = M_{ij}/\mu$ between each pair of coexisting populations.
- $k - 1$ population split times.

**Caveat:** Even for modest values of $k$, there are many parameters that need to be estimated (e.g., 45 when $k = 5$) which is only possible with substantial data and computer time.
Fig. 9: Effect of the prior on the posterior distribution of migration rates
### Table 2. False-Positive Rates and Statistical Power for Detecting Migration.

<table>
<thead>
<tr>
<th>Simulation Model No.</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of loci</td>
<td>10</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>$m$ prior term</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>5.0</td>
<td>$0.05^a$</td>
</tr>
<tr>
<td>False positives $^b$</td>
<td>0.025</td>
<td>0.042</td>
<td>0.075</td>
<td>0.200</td>
<td>0.0130</td>
</tr>
<tr>
<td>$m_1 &gt; 2$ power $^c$</td>
<td>0.1</td>
<td>0.25</td>
<td>0.95</td>
<td>0.667</td>
<td>0.600</td>
</tr>
<tr>
<td>$m_2 &gt; (1,2)$ power $^c$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.70</td>
<td>1.0</td>
<td>0.050</td>
</tr>
</tbody>
</table>