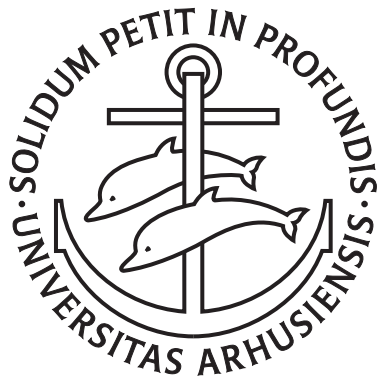
The Coalescent in Structured Populations

Kasper Munch



Department of Genetics and Ecology
Institute of Biological Sciences
The Faculty of Science
University of Aarhus
Denmark

May 23, 2002

Contents

Preface

This thesis is written in order to full fill the requirements of the masters degree. The topic is the retrospective population genetic analysis of structured populations. It is a purely theoretical work, and in that respect it differs somewhat from the tradition, of biological master thesises. I have chosen this form out of fascination of the strength of the retrospective analysis, that is, the Coalescent.

In chapter one a review of the Coalescent is given. This is a probabilistic description the ancestral relationship of sampled sequences. It is in this framework, that the results presented in the thesis, are based. It is assumed throughout that the sequences have not been subject to selection, that recombination of sequences does not occur, and that population size is constant through time.

Chapter two contains a description of the coalescent in a structured population. This is the topic of this thesis, and the chapter serves to introduce a general understanding of the effects of structuring on the ancestral relationship of sampled sequences.

Chapter three covers my own work. Here a structured Moran model is presented to describe a source-sink functionality in a coalescent framework. The model is developed for an island model, and serves to investigate the effects of varying demographic parameters in a structured population of constant size. The Moran model is chosen since this includes the birth and death rates responsible for demographic differences between subpopulations. The approach taken is to resolve all transition probabilities in the structured coalescent into the birth and death rates that produce them. By investigating the coalescence time of two sequences in a source-sink system of two subpopulations, it is shown that the effect of a source-sink dynamic is an effect on effective population size only, if the subpopulation sizes are just moderately large. A result for the source-sink effective population size is presented for the case of strong migration, and the effect on genealogy structure, for small subpopulations, is described.

Chapter four is a discussion of my own work, and of the problems of ambiguity encountered in retrospective genetic analysis.

I would like to thank my supervisor for advice and fruitful discussions during the preparation of this thesis. Further Jakob Skou Pedersen and Roald Forsberg must be thanked for constituting a pleasant working environment.

Chapter 1

The Coalescent

This chapter is concerned with the probabilistic description of the genealogical relationship of sequences sampled from a panmictic population. It is assumed that the sequences an their ancestors have not been subject to selection.

Consider a panmictic population under the Wright-Fisher model. Random sampling governs the representation of lineages through time, and hence, the ancestral relationship of sequences sampled from such a population. An example of how genetic drift affects the representation of lineages in time is shown in figure 1.1. A straight forward consequence of genetic drift is, that the ancestral relationship of the sequences in the present generation can be represented by a tree structure. In figure 1.2 the ancestral relationship of five lineages from figure 1.1 is shown.

Such trees have branch levels, that are characterised by the number of lineages left from the sample, and are separated by events where two lineages find a common ancestor. A probabilistic model, called The Coalescent, that describes this process of ancestral relationship between lineages in a sample was presented by Kingman in his two key papers (Kingman 1982*b*), (Kingman 1982*a*). Coalescent theory has become one of the foremost tools for population geneticists, when making inferences on the trees representing the ancestral relationship of a sample of DNA sequences. It is within this framework, that the models in the remaining part of this thesis are treated. Below the main features of The Coalescent are presented.

1.1 The Coalescence Process

The Coalescent is retrospective, in the sense that is works its way backwards in time describing the ancestral process. When referring to time in the Coalescent it we thus refer to the length of time from the present and backwards in time.

The Kingman Coalescent process, $\mathfrak{A}_t$, is a Markov process in continuous time in which the branch levels are states, and the events where two lineages find a common ancestor, the coalescence events, are transitions between states. A Markov process is a process that has no memory, in the sense that the transition probabilities are only dependent on the state, that the chain is presently in. Hence, the chain of states proceeds from the state where the all the lineages are separate, to the absorbing end state, where all the lineages have coalesced, so that only one

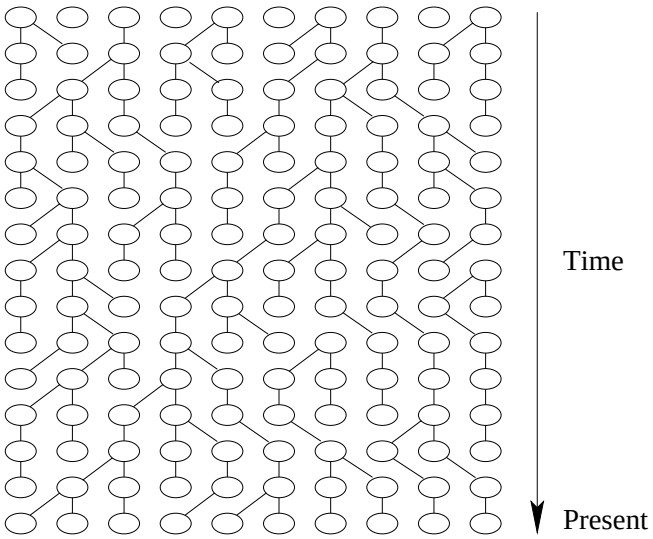


Figure 1.1: As time progresses form the past, the random sampling of gametes to the new generations will give some lineages more descendants at the expense of others. This effect in denoted genetic drift.

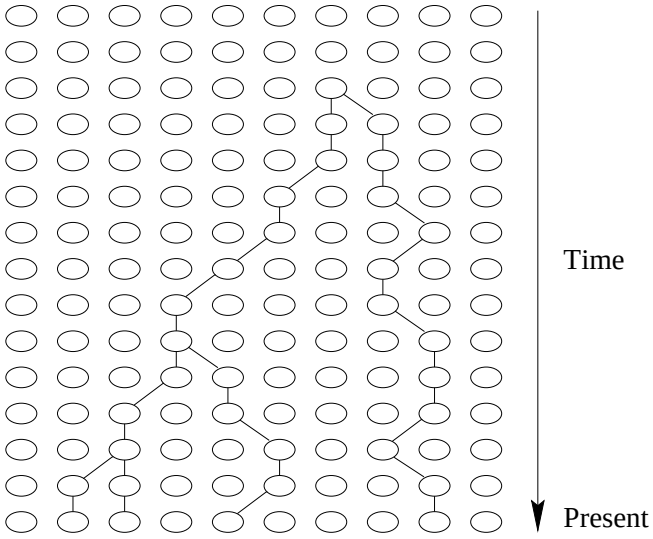


Figure 1.2: This figure is essentially the same is figure 1.1, except that only the only the ancestral relationship of a sample of five sequences from the present population is drawn. As can be seen, owing to genetic drift, the ancestral relationship of a sample takes the form of a tree.

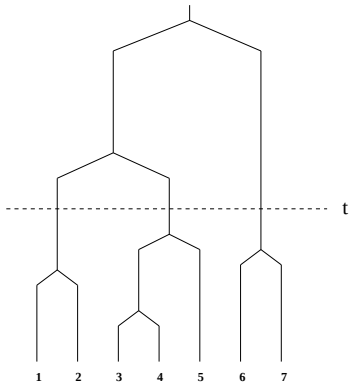


Figure 1.3: The genealogy of a sample of seven sequences. The dashed line indicates time t back in time, and the sampled lineages are by this time represented by only three ancestors or equivalence classes.

lineage is left. This last lineages is denoted the most recent common ancestor. The sequence of states between the initial and the end state, depends on which lineages coalesce. In other words, how the topology of the tree develops backwards in time, is dependent on the way that common ancestors are found among the lineages.

Kingman described this in terms of equivalence relations. An equivalence relation describes how many lineages that are left from the sample at some time in the past, and how many lineages in the original sample that each ancestor is ancestor to. If there are k lineages left at time t, the equivalence relation is a set of k equivalence classes, each representing a lineage at time t. Labeling the originally sampled sequences $\{1 \dots n\}$ one equivalence class contains the labels of the lineages that one ancestor is ancestor to. In other words, an equivalence class corresponds to an ancestor. For $n = 7$, as in figure 1.3, the equivalence relation at time t could be:

$$(1, 2), (3, 4, 5), (6, 7)$$

In this case there are tree lineages left. These are ancestors to sequences 1 and 2, to sequences 3, 4 and 5, and to sequences 6 and 7 respectively. Thus the process moves through a series of equivalence relations with decreasing number of equivalence classes, corresponding to the decreasing number of ancestors. We denote the initial state with n equivalence classes, Δ, and the absorbing end state with one equivalence class ,Υ. The number of equivalence classes at time t is denoted $A_t = |\mathfrak{A}_\mathfrak{t}|$.

The set of equivalence relations with k equivalence classes is denoted Φ_k. It is obviously only possible to get to a state $\eta \in \Phi_k$ from a state $\xi \in \Phi_{k+1}$, and only a subset of Φ_{k+1} will by a coalescence of two equivalence classes produce a particular member of Φ_k. A state ξ that in one step can reach state η is denoted $\xi \prec \eta$ (formally: $\xi \prec \eta = \xi \subset \Phi_{|\xi|},\ |\xi| = |\eta| + 1$).

If the number of genes in the population is N, then the probability that two

particular lineages find a common ancestor in the previous generation is $1/N$. For large values of N it can be assumed that the probability that more than two lineages find a common ancestor is negligible. Thus the probability of transitions between equivalence relations is:

$$p_{\xi\eta} = \delta_{\xi\eta} + r_{\xi\eta}N^{-1} + O(N^{-2}), \tag{1.1}$$

where $\delta_{\xi\eta}$ is the Kronecker delta, (which is one if $\xi = \eta$ and zero otherwise), and

$$r_{\xi\eta} = \begin{cases} -k(k-1)/2 & \text{if } \xi = \eta \text{ and } k = |\xi| \\ 1 & \text{if } \xi \prec \eta \\ 0 & \text{otherwise.} \end{cases} \tag{1.2}$$

$k(k-1)/2 = \binom{k}{2}$ equals the number of possible coalescences between k equivalence classes. If $\mathcal{P} = \{p_{\xi\eta}\}$ is the matrix of these transition probabilities, then scaling time in units of N and passing N to infinity, so that each time step becomes infinitely small, produces the Coalescent:

$$\lim_{N\to\infty} \mathcal{P}^{[Nt]} = e^{Rt}, \tag{1.3}$$

where the superscript $[Nt]$ indicates that time is scaled with N, and where $R = \{r_{\xi\eta}\}$ is the infinitesimal generator of the continuous Markov process $\mathfrak{A}$.

If we for now only consider the process of decreasing the number of ancestors, A_t, it does not matter which equivalence class we have after the transition, but only that we have a transition, so that A_t decreases by one. Since each equivalence relation can produce a new equivalence class by coalescing $k(k-1)/2$ different pairs of equivalence classes, the infinitesimal generator $Q = \{q_{ij}\}$, or the exponential intensities of the continuous Markov process, A_t is:

$$q_{ij} = \begin{cases} k(k-1)/2 & \text{if } j = i-1 \\ -k(k-1)/2 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases} \tag{1.4}$$

A_t only determines the number of equivalence classes (ancestors) at time t. Which of the equivalence classes in a equivalence relation that amalgamate, and thus which of the possible new equivalence relations we have after the transition is determined by another process. This process, governing which of the states with k equivalence classes the process is in, is denoted E_k. E_k is also a Markov process and its transition probabilities are given by:

$$\mathbf{P}(E_{k-1} = \eta \mid E_k = \xi) = \begin{cases} \frac{2}{k(k-1)} & \text{if } \xi \prec \eta \\ 0 & \text{otherwise.} \end{cases} \tag{1.5}$$

Hence, which of the pairs that coalesce, and thus which $\eta \succ \xi$, that is reached from ξ is uniformly distributed.

Since the lineages are indistinguishable, there is no information contained in knowing which of the particular pairs that coalesce. The information is embedded

in the distribution of the sizes of equivalence classes at a given time. Given that we have k lineages remaining in the sample, the probability of reaching a particular equivalence relation is given by:

$$\mathbf{P}(E_k = \xi) = \frac{(n-k)!k!(k-1)!}{n!(n-1)!}\,\rho_1 \ldots \rho_k, \tag{1.6}$$

where $\rho_1 \ldots \rho_k$ are the sizes of the equivalence classes in ξ (Kingman 1982*a*).

E_k and A_t are independent, which is to say that the process that determines the number of lineages remaining from the sample at time t, is independent of the process that determines which lineages coalesce in each event. This means that the Coalescent can be expressed in the form:

$$\mathfrak{A}_\mathfrak{t} = E_{A_t}. \tag{1.7}$$

In words, the probability of having a particular equivalence relation at a particular time, can be factorised into the probability that we at time t have a state with k equivalence classes, times the probability that among all the possible equivalence relations in Φ_k we have a particular one:

$$\mathbf{P}(\mathfrak{A}_t = \xi) = \mathbf{P}(A_t = j \mid A_0 = n)\mathbf{P}(E_k = \xi), \tag{1.8}$$

where $\mathbf{P}(E_k = \xi)$ is given by (1.6), and

$$\mathbf{P}(A_t = j \mid A_0 = n) = \sum_{k=j}^{n} \tau_k^0(t)\frac{(2k-1)(-1)^{k-j}j_{(k-1)}n_{[k]}}{j!(k-j)!i_{(k)}}, \quad 2 \le j \le n, \tag{1.9}$$

where $\tau_k^0(t) = \exp[-k(k-1)t/2]$ (Tavaré 1988). For $n = N = \infty$ (1.9) gives the number of distinct ancestors of the entire population at time t.

As long as the lineages are identical, the results above are of limited value. However, when a Poisson process of mutation on branches of the tree is included, they can be used to calculate the probability, that a mutation shared by some number of sequences in the sample, occured at a particular point in time.

Below only the process of decreasing the number of lineages, A_t, will be considered. The time between two such events, the coalescence time is exponentially distributed. Their mean and variance are:

$$E[T] = \lambda^{-1} \quad \text{and} \quad Var(T) = \lambda^{-2}. \tag{1.10}$$

In deriving the Coalescent under the Wright-Fisher model, we have assumed that multiple coalescence events and coalescence events of three lineages to one ancestor do not happen. This makes the Coalescent an approximation to the Wright-Fisher model. However, with large N, it is a very good one.

Since the expected lengths of the branches are given by $2/k(k-1)$, the expected total length of the tree, $T_{n\to 1}$, for a sample of n lineages, or the time to the most recent common ancestor, of the entire sample is:

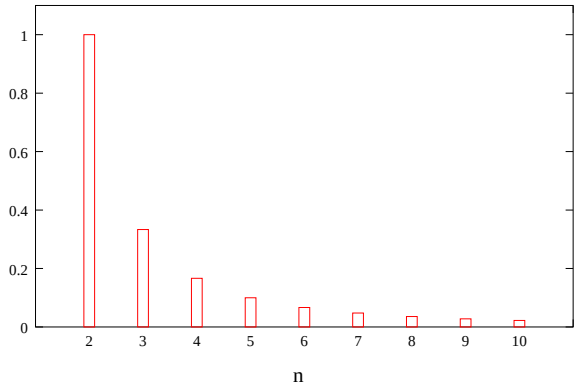


Figure 1.4: The plot shows $E(T_{n\to n-1})$ as a function of n. The last three or four branches constitute together almost the entire length of the tree.

$$E[T_{n\to 1}] = 2\left(1 - \frac{1}{n}\right), \tag{1.11}$$

which is bounded, in that passing n to the limit $n \to \infty$, (1.11) converges to two. (1.11) implies that the time until the coalescence of the last two ancestors, $T_{2\to 1}$, constitutes at least half of $T_{n\to 1}$, and that the expectation of $T_{n\to n-1}$ decreases rapidly as n increases (see figure 1.4). For large n, the first part of the Coalescent process is practically an implosion of lineages. As a result $T_{2\to 1}$ and $T_{3\to 2}$ accounts for most of the variability in $T_{n\to 1}$, see figure 1.4, (Donnelly & Tavaré 1995). This implies, that the time to the most recent common ancestor of a relatively small sample, almost equals that of the entire population.

A nice way of picturing the process of decreasing number of ancestors, is by the densities of having a particular number of ancestors remaining in from the sample. This is depicted in figure 1.5 that in kindly made available by Roald Forsberg and Jotun Hein from a paper of their's to appear.

1.2 Robustness of the Coalescent

Kingman derived the Coalescent for the Wright-Fisher model. But the results are equally valid for other models as well.

If we label all members of a generation, g, $\{1, \ldots, N\}$, then ν_i is the number of offspring in generation $g+1$ to the member with label i in generation g. For the Wright-Fisher model the vector describing the new generation, $\{\nu_1, \ldots, \nu_N\}$ is a symmetric multinomial. The joint distribution of ν_i is said to be exchangeable if it can be assumed, (I) that the members in a generation need not be labelled in any particular way, and (II) that we can assume that the $\nu_i,\ i \in \{1, \ldots, N\}$ are independent of the $\nu_i,\ i \in \{1, \ldots, N\}$ in other generations.

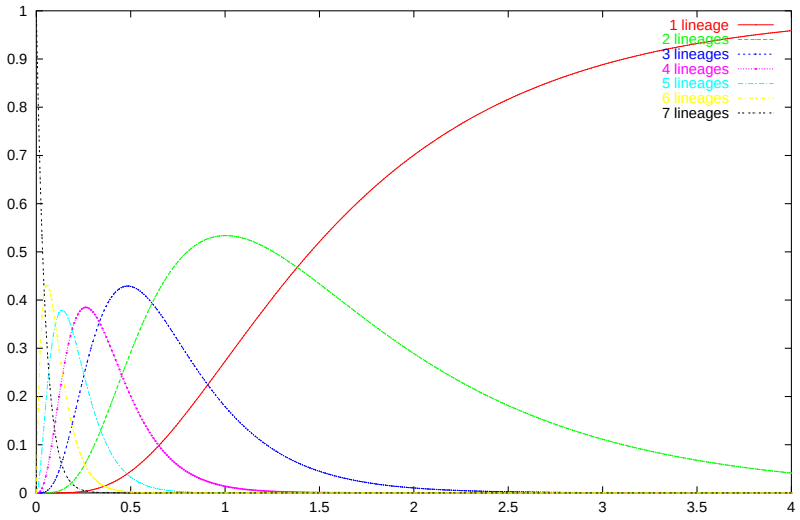


Figure 1.5: The densities of having a particular number of ancestors remaining from a sample of seven sequences.

Kingman showed that the results for the Coalescent as derived for the Wright-Fisher model, with appropriate scaling, are equally valid for any model if the assumptions (I) and (II) applies to the joint distribution of ν_i, and if it can be shown that the variance of the joint distribution of ν_i, σ^2, converges to a finite value as N tends to infinity. (and that the moments of ν are bounded: $E[\nu_i^m] \leq M, \;\; m = 1, 2, \ldots$ for some number M). If this is the case, the results for the Coalescent applies, but with a time scaling N/σ^2. In the Wright-Fisher model $\sigma^2 = 1$. For the Moran model $\sigma^2 = 2/N$ so that the time scaling in this case becomes $N^2/2$.

1.3 The Mutation Process

The mutation process works to differentiate the lineages from the time of their common ancestor and forward to the time of sampling. If this differentiation affects how many offspring each lineage in the population have, or whether or not the lineage is likely to migrate to another population, the genealogy will be dependent on the mutation process. If on the other hand we assume that the mutations are neutral, the mutation process and the genealogical process are independent.

It will be assumed throughout, that the mutational process is a Poisson process with mutation rate μ and mean number of mutations μt. This implies, that the expected number of mutations in a lineage on a time interval t, is linear function of t, and means that the distinction between branch length and number of mutations is a simple matter of scaling.

The waiting time to the next mutation event in a lineage is exponentially dis-

tributed

$$\mathbf{P}(T > t) = \exp\left(-\frac{\theta}{2}t\right), \tag{1.12}$$

where $\theta = 2N\mu$ is a composite parameter that is often used, since μ and N can not be separately estimated on less one of them is known.

The number of mutations separating two sequences is governed by two competing exponentials. Whether two lineages will coalesce first, or whether one of the two will mutate first is governed by the relative size of the coalescence rate and the mutation rate. Since the two processes are independent, the probability that they coalesce first is

$$F = \frac{1}{1+\theta}, \tag{1.13}$$

which is the familiar result for identity by descent. Hence, the expected number of mutations occurring in both lineages before they coalesce, is geometrically distributed with parameter $1/(1+\theta)$.

1.4 Measures of Divergence

The two most widely used measures of divergence between sequences are the number of segregating nucleotide sites, and the average number of pairwise differences in nucleotides. The average number of pairwise differences is the average number of differences between two sequences randomly chosen from the population. This number can be estimated by.

$$E[\Pi_{ij}] = \frac{2}{n(n-1)}\sum_{i\neq j}\Pi_{ij} = \theta, \;\; i,j \in \{1\ldots n\} \tag{1.14}$$

(Tajima 1983), where Π_{ij} is the number of differences between sequences i and j. This follows from the fact that the mean branch length separating two sequences is $2N$, and that the rate of mutation in any of the lineages is 2μ.

The number of segregating sites is the number of sites in the compared locus, in which the sampled sequences differ. Both S and Π have to be scaled with the length of the sequence to obtain a measure useful for comparison measures. However, in the following, when referring to these measures, μ will be the mutation rate per sequence of equal length, so that scaling is not needed. Under the infinite sites model and with random mating, one mutation corresponds to one segregating site. Under the same assumptions the mean number of segregating sites is:

$$E[S] = N\mu\sum_{k=2}^{n} kE(T_{k\to k-1}) = \frac{\theta}{2}\sum_{k=2}^{n} k\,\frac{2}{k(k-1)} = \theta\sum_{k=1}^{n-1}\frac{1}{k} = a\theta, \tag{1.15}$$

where a is the sum $(1+1/2+\cdots+1/n-1)$ (Watterson 1975). Here, the first term of (1.15) is the total branch length, (the mean length of each branch level, times the number of lineages in that branch level), times the mutation rate, scaled with N. Note that the total branch length is given by $\sum_{k=1}^{n-1} 2/k$. So for two sequences the two measures have the same mean.

An advantage of the number of segregating sites over the number of pairwise differences, is that it has a smaller variance, but a drawback is that the number of segregating sites obviously depends on sample size (Li 1997). The number of pairwise differences takes the frequencies into account, whereas the number of segregating sites does.

A way of characterising the proportions of the tree is through the sizes of external and internal branches. A branch is said to be external if it has one end at $t = 0$. Otherwise it is internal. The mean number of segregating sites in external branches is given by:

$$E[S_e] = \theta. \tag{1.16}$$

The mean number of segregating sites in internal branches is:

$$E[S_i] = (a-1)\theta. \tag{1.17}$$

(Fu & Li 1993) The relation between these two measures can be used to test deviations in tree structure, from the expected branch length proportions of the standard Kingman Coalescent.

Chapter 2

The Structured Coalescent

In this chapter, the structured Coalescent is described. The first part contains the general probabilistic description of the genealogical process in structured populations. Only the process of decreasing number of ancestors, denoted A_t i the previous section, will be considered. The second part summerises the general effects of structuring on the Coalescent. The third part is a brief presentation of different migration regimes. In the last part the implications of structure is described in detail in terms of an island model.

For the sake of exhaustiveness I will limit the scope of the following to island models with abstract structure. That is, I will not consider explicit geographical structure such as stepping stone, lattice or torus models. Further I will not consider results for varying deme sizes and hence neither results for meta-population structure. I choose this seemingly narrow scope, because this is intended as a review on the very nature of structuring, and not of the many other effects that is often considered in conjunction with structure. The motivation for doing this, is to review the variety of ancestral relationships, that these relatively simple models can produce.

2.1 The Coalescent with Migration

Here the structured Coalescent will be considered for a haploid species in the island model. However, the results derived account for the diploid setting as well, with population size equal to $N_T/2$, if the following applies: (I) the species is monoecious, (or dioecious if the migration pattern is sex independent). (II) within-population mating is random with selfing in each deme at a rate equal to the reciprocal of the deme size. (III) migration is gamete migration (Nagylaki 1980).

Consider a population subdivided into a number of subpopulations or demes. The set of D demes is denoted $S = \{1, \ldots, D\}$. From the time of sampling and back to the most resent common ancestor, a decreasing number of lineages will at random times change their location among the D demes. This results in a Markov chain with state space I. Each state $\alpha \in I$ is a d-vector that describe the location of the sample among the D demes. So, α_i is the number of lineages in deme i. Such a vector is denoted ε^i if $\alpha_i = 1$ and $\alpha_j = 0$ for $j \neq i$. $|\alpha|$ designates the number

of lineages left in the sample.

2.1.1 Forward migration

When considering forward migration among demes, we must bare in mind, that the object of study is the migration of individuals or gametes. Migration rates are only a modelling tool, to describe this process. Keeping this in mind will ease understanding of the structured Coalescent.

A migration rate is the probability that one particular gene in a deme migrates in one particular generation. It can be interpreted as the number of migrants from deme i to j divided by the size of deme i. That is, the forward migration rate is expressed in terms of the corresponding deme size that genes emigrate from.

The forward migration regime governing the location of the lineage in the population is described by a Markov chain. The transition probability of this Markov chain is the per generation probability, that a lineage migrates from deme j to deme i: m_{ji} for $j \neq i$ and $1 - \sum_{j \neq i} m_{ji} = 1 + m_{jj}$ for $j = i$ where

$$m_{jj} = -\sum_{j \neq i} m_{ji}. \tag{2.1}$$

The model can be deterministic in the sense that a fraction m_{ji} of deme j migrates to deme i each generation, giving $m_{ji}N_j$ migrants from j to i, and it can be stochastic, that is, each individual migrates independently giving a binomially distributed number of migrants, $m_{ji}N_j$.

Nested within these two models, we also have to distinguish between gamete migration and individual migration. With gamete migration the new generation in each deme is sampled partly from the gamete pool of the deme, and partly from the gamete pools of the other demes, or from a gamete pool of migrants from all demes. Here the migration is an integrated part of the sampling process. In such a model migration is more properly denoted dispersal for three good reasons: (I) Migration is not dependent on any properties of the deme. There can be no connection between potential over or under-production and migration, if all gamete pools can be assumed to be infinitely large. (II) Gametes do not think, and they certainly do not evaluate their possible success in a new deme. (III) Gametes spread by external forces independent of habitat quality.

With individual migration, the migration step happens before sampling of the next generation. If the number of immigrants and the number of emigrants are not the same, the deme size is obviously altered, until the sampling to the next generation among the lineages in the deme, regulates the deme size to the original size, N_i. This model allows for a possible evaluation the habitat quality. Hence, individual migration may be caused by a source-sink relationship between demes, and is thus more in line with the conventional notion of the term migration. In the following, however, I will stick to convention anduse the migration term for both gamete and individual migration.

2.1.2 The Genealogical Process

Let $P_t(\beta_i|\alpha_i)$ be the transition probability from α_i to β_i. In the case of gamete migration, the transition probability is straightforwardly given by

$$P_t(\beta_i|\alpha_i) = \begin{cases} \binom{\alpha_i}{2} \Big/ N_i + O((1/N)^2) & \text{if } \beta_i = \alpha_i - 1 \\ 1 - \binom{\alpha_i}{2} \Big/ N_i + O((1/N)^2) & \text{if } \beta_i = \alpha_i \\ O((1/N)^2) & \text{otherwise} \end{cases}, \tag{2.2}$$

since the migration step does not influence the deme sizes.

In the case of individual migration, migration happens before sampling to the next generation, from the deme gamete pool. Because the migration step is separate from the sampling step, the number of different lineages in the gamete pool, is dependent on whether migration causes a net influx or efflux from the deme. With a net influx the number of lineages in the deme before sampling, and thus the number of lineages that constitute the gamete pool, is larger, and the other way around, obviously, for a net efflux. If we let M_{*i} denote the number of lineages added to the deme after migration (M_{*i} may be negative), then the transition probabilities are

$$P_t(\beta_i|\alpha_i) = \begin{cases} \binom{\alpha_i}{2} \Big/ (N_i + M_{*i}) + O((1/N_i + M_{*i})^2) & \text{if } \beta_i = \alpha_i - 1 \\ 1 - \binom{\alpha_i}{2} \Big/ (N_i + M_{*i}) + O((1/(N_i + M_{*i}))^2) & \text{if } \beta_i = \alpha_i \\ O((1/N_i + M_{*i})^2) & \text{otherwise.} \end{cases} \tag{2.3}$$

It is assumed, that the migration rate scales with N_T as N_T goes to infinity, so that the number of migrants stays finite:

$$\lim_{\substack{N_i \to \infty \\ i \in S}} c_i N_T m_{ji} = M_{ji} \tag{2.4}$$

where c_i is the fraction of the total population size, N_T, that deme i constitutes. Scaling time in units of $N_T \to \infty$ so that each time step goes to zero, we get the continuous Markov process for the genealogical process, A_t, with infinitesimal generator for both gamete and individual migration

$$\lim_{\substack{N_i \to \infty \\ i \in S}} N_T\{P_t(\beta|\alpha) - \delta_{\alpha,\beta}\} = \begin{cases} \binom{\alpha}{2} c_i^{-1} & \text{if } \beta_i = \alpha - \varepsilon^i \\ -\binom{\alpha}{2} c_i^{-1} & \text{if } \beta = \alpha \\ 0 & \text{otherwise,} \end{cases} \tag{2.5}$$

The only difference between this result and (1.4)) is, that this is a result for one out of several demes, that together form the population.

In cases where assumption (2.4) do not hold, where the number of migrants M_{*i} can not be assumed to be negligible compared to the deme sizes, the genealogical process under individual migration can not safely be described by the simplification (2.5) unless it is assumed that the deme sizes are regulated to N_i before reproduction.

2.1.3 Backward Migration

The forward migration rates were expressed in terms of the size of the deme, that the lineage was in before the migration event, that is, the deme that the lineage emigrates from.

The backward migration rate is the per generation probability that one particular gene/lineage resident in deme i is received from deme j. In other words, the backwards migration rates are expressed in terms of the deme size, that the lineage ends up in after the forward migration event, rather than the deme size, that the forward migration came from. Since we assume that each migration event does not change the deme sizes, the rates for the backward process is given by

$$r_{ij} = \begin{cases} \frac{N_j}{N_i} m_{ji} & \text{if } i \neq j \\ -\frac{1}{N_i}\left(\sum_{h \neq i} N_h m_{hi}\right) & \text{if } i = j \end{cases}, \tag{2.6}$$

where r_{ij} designates the probability that an individual located in deme i was located in deme j in the previous generation. Hence, just as for the forward rates we have that $\sum_j r_{ij} = 0$ and that $r_{ii} = -\sum_{i \neq j} r_{ij}$. The '$r$' refers to relocation. I will use this word instead of migration, to emphasise that the process of the lineages changing location backwards in time, is not to be thought of as a process of actual migrations happening backwards in time, but only as a set of probabilities, describing migration in a retrospective fashion. The matrix, describing this process, will nevertheless be referred to as the backwards migration matrix.

When modelling population genetics backwards in time, it is crucial to consider throughly which properties of the corresponding forward model, that may inferred from the backwards model, and which features that can not be inferred. This will be addressed in the discussion.

In the Markov chain describing the backwards migration process, time is scaled with N_T, so that it becomes continuous as N_T is passed to infinity. It is assumed that the relocation probability scales with the N_T as $N_T \to \infty$. This implies that the number of migrants do not go to infinity as the population size does. Formally it is assumed that

$$\lim_{\substack{N_i \to \infty \\ i \in S}} c_i N_T r_{ij} = R_{ij}, \tag{2.7}$$

which is a consequence of (2.4). Under the assumption (2.7), we ignore terms smaller than $O(1/N)$. This means that we ignore the probability of having more than one relocation event affecting our sample each generation. The Q-matrix governing the continuous Markov process of relocation events, is then given by

$$\lim_{\substack{N_i\to\infty \\ i\in S}} N_T\{P_t(\beta|\alpha)-\delta_{\alpha,\beta}\} = \begin{cases} \alpha_i r_{ij} c_i N_T & \text{if } \beta = \alpha - \varepsilon^i + \varepsilon^j \text{ for } i \neq j \\ -\sum_{i\in S} \alpha_i |r_{ii}| c_i N_T & \text{if } \beta = \alpha \\ 0 & \text{otherwise,} \end{cases} \tag{2.8}$$

since the individual lineages change location independently. Again the time is scaled in units of N_T as N_T is passed to infinity. The Markov process governed by the infinitesimal generator (2.8), can also be thought of as a system of $|\alpha|$ Markov processes, each describing the location of one particular lineage.

2.1.4 The Combined Process

Combining the independent Markov processes of the migration and the genealogy, we get the continuous Markov process for the structured Coalescent. The infinitesimal generator of this Markov chain is the matrix:

$$Q_{\alpha,\beta} = \begin{cases} \alpha_i r_{ij} N_T & \text{if } \beta = \alpha - \varepsilon^i + \varepsilon^j \text{ for } i \neq j \\ \binom{\alpha_i}{2} c_i^{-1} & \text{if } \beta = \alpha - \varepsilon^i \\ -\left\{ \sum_{i\in S} \binom{\alpha_i}{2} c_i^{-1} + \sum_{i\in S} \alpha_i r_{ij} N_T \right\} & \text{if } \beta = \alpha \\ 0 & \text{otherwise.} \end{cases} \tag{2.9}$$

It is implicit that the infinitesimal generator of the combined process is only given by Q if $|\alpha| \geq 2$ and is zero otherwise. That is, any state $\alpha = \{\alpha \in I;\ |\alpha| = 1\}$ is an absorbing state. The waiting time to the next event of any kind is thus exponentially distributed, with the proper diagonal entry in the Q-matrix as parameter.

$$\mathbf{P}(T < t) = 1 - \exp\left[-\left(\sum_{i\in S} \binom{\alpha_i}{2} c_i^{-1} + \sum_{i\in S} \alpha_i R_{ij} \right) t \right], \tag{2.10}$$

where $R = c_i N_T r_{ij}$ is the scaled relocation probability. Since the time to the next event of each particular type is independent, the process can be resolved into $2 \times D$ competing exponentials (a migration or a coalescence event for each deme), each with an expected waiting time equal to the reciprocal of its intensity. With N_T going to infinity, the probability of more than one event of either coalescence or relocation is negligible, since both coalescence rate and relocation rate is $O(1/N_T)$.

When an event of any type occurs, the probability that it is a particular type event, e.g. a coalescence event in deme two, is a simple weighting of the exponential intensities:

$$\mathbf{P}(\text{Next event = Coal. in deme two}) = \frac{\binom{\alpha_2}{2} c_2^{-1}}{\sum_{i \in S} \binom{\alpha_i}{2} c_i^{-1} + \sum_{i \in S} \alpha_i R_{ij}}. \qquad (2.11)$$

Results such as the probability that some number of lineages coalesce in one deme before any of them are relocated to another deme is of cause straightforwardly obtained from this property of exponential distributions. Slatkin's (1989) result for non-immigrant ancestry, the probability that all the α_i lineages sampled in a deme coalesce before any of them relocates to another deme exemplifies the applications. Here, even deme sizes are assumed for simplicity.

$$\mathbf{P}(n) = \prod_{k=2}^{\alpha_i} \frac{\binom{k}{2}}{\binom{k}{2} + k r_{i*} c_i N_T} = \prod_{k=2}^{\alpha_i} \frac{k(k-1)}{k(k-1) + k 2 R_{i*}} = \frac{(k-1)!}{\mathcal{B}_{(\alpha_i)}}, \qquad (2.12)$$

where the subscript $i*$ means deme i to some other deme, and $\mathcal{B}_{(\alpha_i)} = (2R_{i*} + 1) \cdots (2R_{i*} + \alpha_i - 1)$. Slightly modified from (Slatkin 1989).

2.1.5 The Structured Coalescent and $\mathbf{F_{ST}}$

Identity by descent results and Coalescent results are equivalent since they are just alternative ways to describe the same ancestral processes. This implies that existing identity by descent results relatively easily can be converted to coalescence results and vice versa. Wright's F_{ST} is defined as.

$$F_{ST} = \frac{f_0 - \overline{f}}{1 - \overline{f}}, \qquad (2.13)$$

where f_0 is the probability of identity of two sequences sampled from the same deme, and $\overline{f}$ is the probability of identity of two sequences sampled sampled at random from the collection of demes. Much work on structured populations is done in terms of F_{ST}. To evaluate these in the newer Coalescent framework, F_{ST} must be expressed in terms of coalescence times. An advantage of the approach is that drift, migration and mutation are independent processes under the Coalescent.

If we assume a Poisson process of mutation, the probability of identity by descent of two sequences, is simply the probability that none of them were subject to a mutation before the time where the pair coalesced. This probability is $e^{-\theta t}$, where θ is the scaled mutation rate $4N\mu$, and t is time. If we let T_0 and T denote the time to the coalescence of two sequences sampled from the same and sampled randomly from the collection of demes, respectively, then

$$f_0 = E[e^{-\theta T_0}], \tag{2.14}$$

(Hudson 1990). Analogously

$$\overline{f} = E[e^{-\theta T}]. \tag{2.15}$$

Thus, using (2.13) we obtain

$$F_{ST} = \frac{E[e^{-\theta T}] - E[e^{-\theta T_0}]}{1 - E[e^{-\theta T}]}, \tag{2.16}$$

(Wilkinson-Herbots 1998). This is the exact F_{ST} value. Slatkin (1991) gave an approximation to the exact result, which is actually (2.16) in the limit $\theta \to 0$. Using l'Hôpitals rule on (2.16) we get

$$\widehat{F}_{ST} = \lim_{\theta \to 0} F_{ST} = \frac{E[T] - E[T_0]}{E[T]}. \tag{2.17}$$

F_{ST} measures based on coalescence times do not make full use of the information provided by DNA sequences. It only uses the information on coalescence times for sequences pairs, and not the information on tree structure also contained in sequence data. Nevertheless, it is used extensively in the studies on population structure.

2.2 General Effects of Structuring

Before we embark on the results obtained for structured populations, an intuitive understanding of the effects of structuring a population, may be a valuable tool in understanding what follows. Generally, structuring of a population results in three different effects, that may produce a Coalescent deviating from the standard Kingman Coalescent:

- *Coalescence Preclusion:* The relocation events in the history of the sample or the mode of sampling, may locate lineages among demes so that some pairs of lineages are not able to coalesce. That the number of pairs of lineages that can potentially coalesce is reduced, obviously reduces the overall coalescence rate. When only one lineage remain in each separate deme, no coalescences can occur, until relocations bring pairs of lineages into the same deme. The time elapsing where all lineages are precluded from coalescing will be denoted "relocation waiting time".

- *Early coalescences/Aggregation:* Relocation events or the lack of relocation events may leave large parts of the sample in a subsection of the population. This will result in a larger coalescence rate, since lineages from the sample will then constitute a larger part of the total number of lineages in the subsection concerned. This may be the case if large parts of the sample is

taken from one deme. In this case several pairs of lineages have an elevated probability of coalescing before the sample spread other demes. This effect is denoted "early coalescences". A similar effect arises if relocation is asymmetric in a way that makes lineages collect in one or a few demes. In this case the coalescence rate is also elevated. This effect is denoted "aggregation of lineages"

- *Local Drift Difference:* When lineages are relocated to other demes, the genetic drift regime imposed on the them is most likely changed. This is obviously the case when a lineage is relocated to a deme of different size than that of the one it was located in before. In addition, as will be addressed in chapter 3, different demographical regimes between demes, may result in local drift differences

These effects will be addressed as they appear in the following.

2.3 General Migration Regimes

In this section I will consider the four basic migration scenarios. This section is intended as a help to grasp the biological implications of the assumptions in backward models. The four scenarios are combinations of two properties of relocation, Namely, whether it is isotropic and whether it is conservative. Migration is denoted isotropic, if the relocation probabilities are the same for all demes. It is conservative, if the number of individuals relocating into a deme is equal to the number that relocates out of the same deme. This assumption obviously have the same implications whether it is expressed in terms of the forward or the backward model. Formally, in terms of the backward model, we must have that

$$N_i \sum_{j \neq i} r_{ij} = \sum_{j \neq i} r_{ji} N_j \tag{2.18}$$

or

$$\sum_{j} r_{ji} N_j - N_i = 0 \tag{2.19}$$

The four general migration scenarios are:

1. *Isotropic and conservative*: As seen from (2.18), this implies that all deme sizes are equal. That both relocation rates and deme sizes are equal for all demes implies that the forward and the backward migration matrix are the same. Hence, in the forward model, the same number of individuals/gametes emigrate from each deme, with an even probability of ending up in any of the other demes. That is, the number of immigrants and emigrants are not just the same for each deme, it is also the same between demes.

2. *Isotropic and non-conservative*: Using (2.18) implies, that if the backwards migration matrix is isotropic, migration can not be non-conservative unless

the deme sizes are uneven. Since migration rates and relocation probabilities do not refer to the same deme, this implies that the forward migration rates are not the same among demes, even though the relocation probabilities are. Hence, in terms of the forward setting, we have a set of demes with uneven migration rates and either a net influx or net efflux of individuals/gametes in each deme.

3. *Non-isotropic and conservative*: If the migration rates are not the same between demes, nor are the deme sizes, if migration is to be conservative. Note that conservative migration only means that the net influx and the net efflux from each deme is the same, not that the number of migrants exchanged, is the same for all demes, or that the exchange of individuals/gametes between pairs of demes is symmetric.

4. *Non-isotropic and non-conservative*: Nothing general can be said about these settings, besides that they are not included in the cases described above. A special case of this scenario is symmetric forward migration rates.

2.4 The Coalescent in an Island Model

In this section, structuring in the island model will be considered. In the island model, no physical distance between demes is involved. However, this does not mean that demes separated by varying physical distance can not be modelled. If larger physical distance is assumed to lower migration probability, features of non-abstract structure can be included in the model, by setting the relocation probabilities between more distant demes to smaller values.

2.4.1 Two Sequences

This section will deal with some special features of the simple two-sequence case. As already mentioned, comparing results for branch length and number of segregating sites is a simple matter of scaling, if a Poisson process of mutation is assumed. For reference, recall that the number of differences between two sequences sampled from an unstructured panmictic population is given by

$$E[S] = 2N_e\mu, \tag{2.20}$$

(Kimura 1969)

$$Var[S] = (2N_e\mu)^2, \tag{2.21}$$

where μ is the mutation rate per DNA sequence, and $N_e = N_T$ is the effective population size of haploid individuals. In coalescence terms, (2.20) follows from the fact that the expected coalescence time of the to sequences in this case is N_e, so that the total branch length separating the two sequences is $2N_e$, and from assuming

a Poisson process of mutation. Hence, the expected coalescence time is obtained by scaling with 2μ, which is the rate of mutation in both lineages. Scaled with N_e the mean and the variance of the coalescence time are both equal to one. In the following, results for expected coalescence time, and for number of segregating sites and total branch length will be listed in parallel, since they are so readily convertible.

Lets now turn to a structured population. Notohara (1990) gave the following general results for the expected coalescence time for two sequences in a system of two demes.

$$E[T|(2,0)] = \frac{c_1(3r_1 + r_2) + 4c_1c_2N_T(r_1 + r_2)^2}{(r_1 + r_2) + 4(c_1N_Tr_1^2 + c_2N_Tr_2^2)} \tag{2.22}$$

$$E[T|(1,1)] = \frac{c_1(2r_1 + r_2) + c_2(r_1 + 2r_2) + 4c_1c_2N_T(r_1 + r_2)^2 + 1}{(r_1 + r_2) + 4(c_1N_Tr_1^2 + c_2N_Tr_2^2)} \tag{2.23}$$

$$E[T|(2,0)] = \frac{c_2(r_1 + 3r_2) + 4c_1c_2N_T(r_1 + r_2)^2}{(r_1 + r_2) + 4(c_1N_Tr_1^2 + c_2N_Tr_2^2)} \tag{2.24}$$

(2,0) denotes sampling from two genes from deme one and zero from deme two. (Takahata (1988) gave a general but even more complicated result for D demes.) These results are only included to show the complexity of general analytical results even for very simple systems. Below some simpler special cases will be considered.

If migration is both isotropic and conservative,implying that all deme sizes and relocation probabilities are even, the results reduces to $E[T|(2,0)] = E[T|(0,2)] = 1$ and $E[T|(1,1)] = 1 + 1/2R$, where $R = N_Tr$. This was originally obtained by Li (1976) who showed, that in a system of D demes the number of differences separating two sequences sampled from the same deme, and from different demes is given by

$$E[S^{(s)}] = 2DN\mu \tag{2.25}$$

$$E[S^{(d)}] = 2DN\mu + (D-1)\frac{\mu}{r}, \tag{2.26}$$

where $N_e = DN = N_T$, if each deme is panmictic. The subscripts signifies whether the two sequences in question are sampled from the same or from two randomly chosen different demes. Rescaling as above, (2.25) and (2.26) become

$$E[T^{(s)}] = 1 \tag{2.27}$$

and

$$E[T^{(d)}] = 1 + \frac{D-1}{2R} \tag{2.28}$$

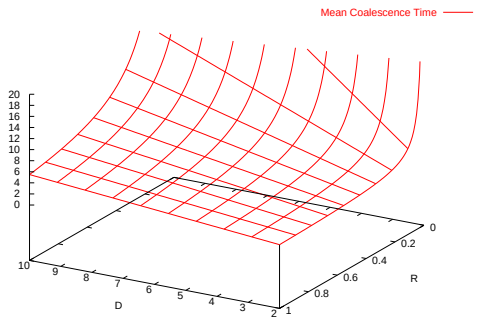


Figure 2.1: The dependence of the expected coalescence time of two sequences on the number of demes, D and the scaled relocation probability, R, when sampling the two sequences from different demes.

respectively, (Notohara 1990), (Hudson 1990) and (Hey 1991). When sampling from different demes, the dependency on structure is straightforward. The lower the relocation probability, the longer the relocation waiting time. As shown in figure 2.1 this effect is stronger the larger the number of demes, since this reduces the probability that the two lineages find each other in the same deme. Note also, that the effect of a higher relocation probability is stronger in a more subdivided population. In contrast, when sampling from the same deme, the mean coalescence time is independent of the backward migration matrix and the number of demes the population is subdivided into. It seems counterintuitive, that relocation probability would not play a role, and surely is does, but the effect of early coalescences and the effect of migration waiting time cancels out, so that no effect of deme sizes or relocation probabilities is seen in the mean coalescence time. The variance of coalescence time, however, shows a dependency on D and R for both both modes of sampling:

$$Var(T^{(s)}) = 1 + \frac{D-1}{DR} \tag{2.29}$$

$$Var(T^{(d)}) = 1 + \frac{D-1}{DR} + \frac{1}{4R^2}. \tag{2.30}$$

This is straightforwardly obtained from the result of Hey (1991). The dependence of (2.29) on relocation probability is intuitively obvious. A lower relocation probability will result in a stronger affect of both early coalescences and relocation waiting time. This implies, that the probability of short and very long coalescence times will increase, thereby increasing the variance.

Slatkin (1987) elaborated on the result by Li, by showing that (2.25) also holds, if migration is only isotropic and not conservative. In this case $E[S_s]$ is calculated, by weighting each deme by the reciprocal of the deme size, so that $E[S^{(s)}]$ is a weighted average of $S^{(s)}$ over all demes,

$$E[S^{(s)}] = \frac{D}{\sum_{i \in S} 1/N_i} \sum_i (S_i^{(s)}/N_i)/D = 2N^{(h)} D\mu, \tag{2.31}$$

where $N^{(h)}$ designates the harmonic mean of the deme sizes, D is the number of demes, and $S_i^{(s)}$ the expected number of segregating sites between two sequences sampled from deme i. Here $N_e = DN^{(h)}$. Scaling with $2DN\mu$ The expected coalescence time is obtained

$$E[T^{(s)}] = \frac{N^{(h)}}{N} = \frac{D}{N \sum_{i \in S} 1/N_i} = \frac{D}{D^{-1} N_T \sum_{i \in S} 1/c_i N_T} = \frac{D^2}{\sum_{i \in S} c_i^{-1}}, \tag{2.32}$$

where N denotes the arithmetic mean of the deme sizes. Note that (2.32) is a result for sequences both sampled in a randomly chosen deme, and not a result applicable to any particular deme, as is (2.27). On the contrary, the expected number of differences for two sequences sampled in some particular deme, is indeed expected to be dependent on migration. Since migration is isotropic, the relocation waiting time when the first lineage relocates from the sampling deme is the same for all demes. On the contrary, the relation between coalescence rate and relocation probability is not the same for demes of different size. This implies that the effects of early coalescences and relocation waiting time will not be the same among demes, and will thus only averagely cancel out, as indicated by (2.32). Two sequences sampled from a small deme, will have a shorter expected coalescence time, whereas sequences sampled in large demes will have a shorter one. Only when $c_i = c$ for any i, the result is equally valid for the collection of demes, and each particular deme. (2.32) shows that if the sizes of the demes are not the same, the expected number of differences, will indeed be affected by structuring, even though the effect is obviously still independent of the backward migration matrix. This follows from the fact that the harmonic mean is always smaller than or equal to the arithmetic mean. Hence, with isotropic migration, the expected coalescence time for two sequences sampled in the same deme, will always be lower than one, except if $c_i = c$ for all i, in which case the means are equal. In this situation migration is isotropic and conservative, and (2.32) collapses into (2.27).

Strobeck (1987) showed that (2.25) is also obtained, with the assumption of week evolutionary forces, as described above for the structured Coalescent. That is, the probability that two events per generation is negligible, be that relocation events, two mutation events, or a combination of both. Under this simplifying assumption he showed that the average expected number of differences is independent of the backward migration matrix, in the case where migration is conservative, but not necessarily isotropic. In this case, the number of segregating sites is given by

$$E[S^{(s)}] = \sum_i N_i S_i^{(s)}/N_T = \sum_{i \in S} c_i S_i^{(s)} = 2N_T\mu. \tag{2.33}$$

Comparing (2.25), (2.31) and (2.33) implies that the condition that makes N_e and thus the expected coalescence time independent of both deme sizes *and* the backwards migration matrix is not whether migration is isotropic, which is the premise shared by (2.25) and (2.31), but rather whether it is conservative, which is the premise shared by (2.25) and (2.33).

Recall the interconnection between F_{ST} and expected coalescence time described in section 2.1.5. A result for F_{ST} can be obtained using the Laplace transforms of the distributions of $T^{(s)}$ and $T^{(d)}$. T^*, the expected coalescence time of two sequences sampled at random among all the demes, is obtained through the expectations of $T^{(s)}$ and $T^{(d)}$. Hence for isotropic and conservative migration F_{ST} is given by

$$F_{ST} = \frac{1}{1 + 2RD^2/(D-1)^2 + \theta D/(D-1)}. \tag{2.34}$$

This approach is due to Wilkinson-Herbots (1998). Slatkin's (1991) approximate result for the finite island model is

$$\widehat{F}_{ST} = \frac{1}{1 + 2RD^2/(D-1)}. \tag{2.35}$$

2.4.2 More than Two Sequences

For samples of more than two sequences, the effects of structuring are still basicly the same. However, the spatial distribution of the sample is no longer necessarily the simple "together or apart" making the scaled coalescence rate in each deme $1/c_i$ or zero. Hence, the coalescence preclusion effect of structuring is no longer entirely a relocation waiting time effect. Rather, the coalescence rate now depends on the number of pairs located in the same demes so that they can potentially coalesce, $\sum_{i \in S} \binom{\alpha_i}{2}$, as well as the sizes of these demes. Singletons obviously have a particularly strong effect on coalescence rate, since these are precluded from coalescing with any other lineage, and in effect does not contribute to the coalescence rate at all. Further, the effect of structuring is, apart from migration rates and deme sizes, dependent on the sample size relative to the number of demes. The more demes, the more strongly the sample may be separated, and the fewer the lineages the fewer demes it takes to separate them. This is the effect depicted in figure 2.1.

Dependence on Sampling

The dependency of sampling is another way of saying that it matters what initial position the sample is in at time zero. The larger the relocation probabilities the less is this dependency. In the limit with infinitely strong migration, there is no dependence on sampling. This special case is considered in section 2.4.4.

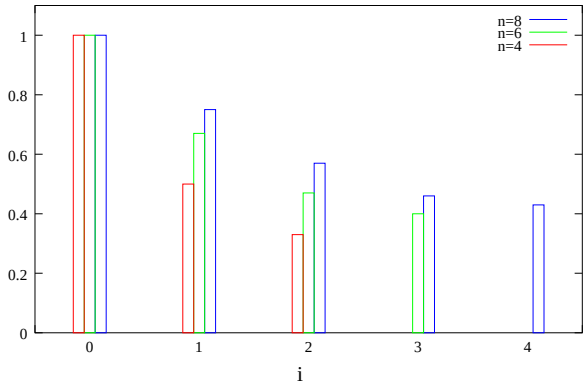


Figure 2.2: For two demes, the figure shows the coalescence rate of n remaining lineages, if $n-i$ is located in one deme and i lineages i the other. The rates are normalised with the coalescence rate when all lineages are in one deme. $n = 4,\ 6$ and 8.

If we assume even deme sizes and even drift regimes in all demes, the effect of some mode of sampling depends on the extent to which the stage is set for either coalescence preclusion or early coalescences.

The number of lineage pairs that are not precluded from coalescing is, in the case of two demes, given by $i(i-1)+(n-i)(n-i-1)$, where n is the total sample size and i is the number of sequences sampled from one of the demes. Hence, if λ_s^0 denotes the instantaneous coalescence rate at the time of sampling for a structured sample, and λ^0 denotes this rate when sampling all sequences from the same deme, the dependence on sampling can be expressed as the fraction

$$\lambda_s^0/\lambda^0 = \frac{i(i-1)+(n-i)(n-i-1)}{n(n-1)}. \tag{2.36}$$

In figure 2.2 equation (2.36) is plotted as a function of i for different sample sizes. Note how the implications of structuring are larger for smaller samples. This is because singletons, and the stronger effect these have on coalescence preclusion, are more probable for smaller samples.

Early coalescences are a result of low relocation probabilities relative to the coalescence rate. Hence the effect may result from both low relocation probabilities, and from sampling several sequences from the same deme. With a lot of demes and a low relocation probability, sampling of the sequences from the same or a few demes, may greatly diminish the time to the most recent common ancestor. This is because most or all lineages from each deme will coalesce before they are spread out into solitude, and thus imposed by the extra waiting time until a relocation into an occupied deme. Recall that for a sample taken from only one deme, the probability that all of the lineages coalesce before any of them migrate is $\prod_{i=2}^{n} \binom{i}{2} / \left(\binom{i}{2} + i r_{ij} N \right)$ (Slatkin 1989). That this is a sum of weighted exponential intensities gives a good perception of the interplay between early coalescences and relocation rate.

Tree Topology

The tree describing the ancestry of the sample can take $n!(n-1)!/2^{(n-1)}$ distinguishable topologies, and the effects of sampling is manifested in the topology of the tree.

Early coalescences will result in monophyletic trees, that is, trees where sequences sampled from the same deme, coalesce to one ancestor for each deme, before any relocation events occur. The larger the relocation probabilities and the more distributed the sample, the more probable will para- or polyphyletic trees become. Hence, given the mode of sampling the topology of the tree contains some information about the migration rates. This information was incorporated into a cladistic measure of migration by Slatkin and Maddison (1989). For a sample from two demes each topology may be characterised by a minimal number of relocation events needed. The approach is a simulation based one, building a catalogue of these characteristic minimal relocation events for an array of relocation probabilities.

Takahata and Slatkin (1989) have studied, under what conditions the three different phylogeny types will result. The probabilities are obtained recursively. With the assumption that $r_{ij} = r$ for all i and j the probabilities of mono- and paraphyly when sampling two sequences from one deme and one from the other, are

$$\mathbf{P}(mono) = \frac{1 + 7R/6 + R^2/3}{1 + 5R/2 + R^2} \tag{2.37}$$

$$\mathbf{P}(para) = 1 - \mathbf{P}(mono), \tag{2.38}$$

since we can not have polyphyly with three sequences. $R = N_T r$. According to intuition $\mathbf{P}(mono) \to 1$ and $\mathbf{P}(para) \to 0$ as $R \to 0$ The results for two sequences sampled from each deme are not given here, but their graphical representation in figure 2.3 give a good perception of the dependency on R. The probability of monophyly is high if the expected number of migrants from each deme is smaller than one.

Their approach, however, is not feasible for arbitrary sample size because the number of Markov states quickly becomes to large to handle. As an approximation it can be assumed that there will be at most one migration event before all demes in each deme have found a common ancestor. Under that approximation the probability of monophyly of a sample taken from two demes obviously correspond to the product of the probabilities for non-immigrant ancestry (see equation (2.12)) for each deme (Slatkin & Maddison 1989).

$$\mathbf{P}(mono) = \mathbf{P}(\alpha_1)\mathbf{P}(\alpha_2) = \prod_{k=2}^{\alpha_1} \frac{k(k-1)}{k(k-1) + kR} \times \prod_{j=2}^{\alpha_2} \frac{j(j-1)}{j(j-1) + jR} \tag{2.39}$$

where α_1 sequences are taken from deme one and α_2 sequences are taken from

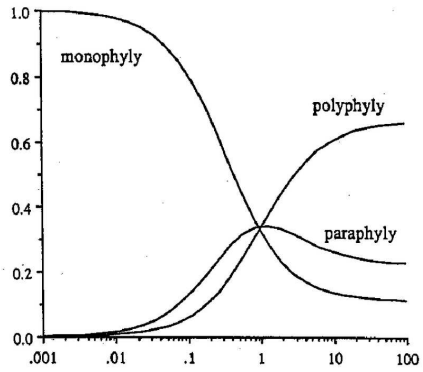


Figure 2.3: The probabilities of monophyly, paraphyly and polyphyly as a function of R for a sample of four genes. Two sequences are sampled from each of two demes. Migration is isotropic. Takahata and Slatkin (1989)

deme two. The probability of paraphyly is still $\mathbf{P}(para) \approx 1 - \mathbf{P}(mono)$, since under the assumption of low levels of migration $\mathbf{P}(poly) \approx 0$.

Expected Coalescence Time

As the number of sampled sequences grows it becomes increasingly difficult to find the expected coalescence time analyticly. The number of Markov states rapidly becomes immense, and so does the number of linear equations, that must be solved simultaneously, to get an exact solution for the mean.

Following the recursive approach of Tajima (1989) Notohara (1990), and Wakeley (1998), stating the process as a Markov chain with X states and infinitesimal generator Q, it is possible to calculate the expected time to the most recent common ancestor, or to any other branch level in the tree as

$$E[T_i] = \frac{1}{q_{i*}} + \sum_{j=1, j\neq i}^{X} \frac{q_{ij}}{q_{i*}} E[T_j]. \tag{2.40}$$

Takahata (1988) approached the problem in essentially the same way. He showed, that for two demes of even size, and with isotropic and conservative migration, the expected coalescence time from three to two sequences is given by

$$E[T(2,1)] = \frac{3+2R}{6+6R} \tag{2.41}$$

$$E[T(3,0)] = \frac{1+2R}{6+6R} \tag{2.42}$$

As $R \to 0$, $E[T(2,1)] \to \frac{1}{2}$ and $E[T(3,0)] \to \frac{1}{6}$. As $R \to \infty$, $E[T(2,1)]$ and $E[T(3,0)] \to \frac{1}{3}$, which is equal to the expected coalescence time of three sequences, $1/\binom{3}{2}$, in the standard Kingman Coalescent. The Coalescent under

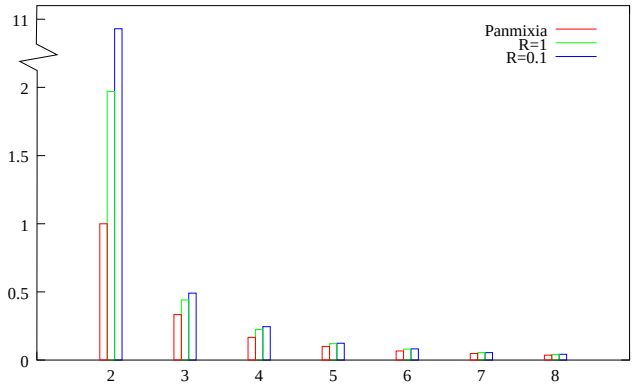


Figure 2.4: The expected coalescence time as a function of the number of lineages in the branch level. The initial state for each branch level is $([n/2], n - [n/2])$, where $[n/2]$ denotes the integer part of $n/2$. The extreme value of 10.93 in the last branch level, corresponding to $R = 0.1$ owes to the relocation waiting time before the lineages become located in the same deme. This is expected to be $1/R$

this strong migration limit is returned to in section 2.4.4. Note that, for small R, the coalescence time for three sequences sampled in the same deme, is smaller than is expected from the Kingman Coalescent. This results from the fact that sampling leaves the lineages aggregated in a subsection of the population, so that a small R results in a strong early coalescence effect. In the limit $R \to 0$ the expectations correspond to the time to the first coalescence of two and three sequences in a population of size $N_T/2$.

Generally the fewer lineages there are to distribute among some number of demes, the stronger the effect of coalescence preclusion. Hence, the effect of coalescence preclusion is stronger in the last part of the tree. This will prolong the last branch levels and thus result in an alteration in the relative proportions of branch levels, compared to the standard Kingman Coalescent. For a system of many demes the effect will not be so pronounced, since the sample in this case most probably will be highly distributed also in the first branch levels. This means that the prolonging effect will be strong in all branch levels, and that the change in the relative proportions of the branch levels will be small. When the number of demes becomes very large relative to the sample size, the relative proportions of the the standard Kingman Coalescent are obtained for the branch levels not negligibly short. This case is considered in section 2.4.5.

In these complicated matters simulation quickly becomes an appealing alternative. Takahata (1988) simulated the mean time between coalescence events, for two demes of equal size. He addressed only the cases where the lineages are evenly distributed between the two demes at the beginning of the branch level. Some of his results are shown in figure 2.4. This is only a part of the general picture, but baring this in mind, it nevertheless presents some general features and the magnitude of the effect of structuring in the different branch levels. The effect of is generally

most conspicuous in the last one or two branch levels. If the sample is not taken evenly among demes, as in figure 2.4, the effect of longer last branches will be even more distinct, since early coalescences will make the first branch levels shorter.

As a shortcut from simulations in these complicated matters, approximations are of great value. Takahata (1991) investigated the Coalescent under the low migration limit for a set of equal sized demes and isotropic migration. In this case it is posited, that the relocation probability is very low relative to the coalescence rate within demes. Hence, if the sample is taken from d of the D demes, it can be assumed, that the time it takes for all the sequences, to find one common ancestor in each deme, is very small compared to the time it takes for the remaining lineages to find a common ancestor. The latter time is much longer, since the relocation events bringing two singletons together in a deme are very rare. When this does happen, it is additionally assumed, that the probability of a coalescence before one of the lineages relocates from the deme again is one. This assumption is valid since it is assumed that the scaled relocation probability is much smaller than one. In conclusion, if the time to find a common ancestor in each deme is negligible compared to the time it takes for the last d lineages to find a common ancestor for the entire sample, the time to the most recent common ancestor of the entire sample is approximated by the time it takes for the last d singletons to find a common ancestor. The mean time to the most recent common ancestor of this simpler process is given by

$$E_{Low}[T_T] = \frac{D-1}{R}\left(1 - \frac{1}{d}\right), \tag{2.43}$$

where $R = DNr$, and the subscript, T, signifies total expected coalescence time. Under these conditions, d is obviously an important parameter in determining the total expected coalescence time. On the contrary, it is only weakly dependent on D.

If, on the other hand, the scaled relocation probability is very large, there is no dependence of sampling, and the total expected coalescence time is approximated by that of a panmictic population (The strong migration limit will be described in section 2.4.4)

$$E_{High}[T_T] = 2\left(1 - \frac{1}{n}\right). \tag{2.44}$$

Takahata showed through simulations, that the low and the high migration approximation is precise for $4Nr \leq 0.1$ and ≥ 10 respectively. He further suggested an interpolation of the two results, to cover the intermediate parameter range

$$E_{Interpol.}[T_T] = \frac{(d-1)(D-1)}{dR}\, 2\left(1 - \frac{1}{n}\right) \tag{2.45}$$

For the appropriate magnitude of migration each of these are good approximations. However, it is obviously a problem that previous knowledge of at least the magnitude of migration is needed to pick the right approximation.

Total branch Length

With a sample larger than two, the conversion between coalescence time and total branch length or number of segregating sites, is not possible unless we know the branch level lengths of the tree. Hence, it is more straight forward to calculate it directly, as described above for expected coalescence time:

$$E[T_{(i)}] = \frac{k}{q_{i*}} + \sum_{j=1, j\neq i}^{X} \frac{q_{ij}}{q_{i*}} E[T_{(j)}]. \tag{2.46}$$

Note that the nominator is k in the first term. k denotes the number of lineages left from the sample. By this approach Wakeley (1998) presented results for the total branch length of trees from samples of three sequences assuming isotropic and conservative migration. As a reference, recall that the total branch length for three sequences in a panmictic population of size $N_T = DN$ is three.

$$E[T_B(3,0,0)] = 3 \tag{2.47}$$

$$E[T_B(2,1,0)] = 3 + \frac{D-1}{2DNm} \tag{2.48}$$

$$E[T_B(1,1,1)] = 3 + 3\,\frac{D-1}{2DNm}, \tag{2.49}$$

where the subscript, B, signifies total branch length. As for two lineages, the expected total branch length, and thus the number of segregating sites in a sample of three sequences, is independent of migration, if the sample is taken from one deme. In other words, it is not possible to to make inferences on the level of of structure from the mean number of segregating sites from this type of sample, as it is not for two sequences from the same deme.

The results (2.48) and (2.49), show an expected dependence on sampling. The more distributed the sample is among demes, the stronger is the dependency on migration, since the possibility for early coalescences decrease as the sample is taken from more demes. The mean total branch length for four sequences, sampled in the same deme, is not independent of the backwards migration matrix (Wakeley 1998).

Based on the results for one, two, three, four and five sequences, Wakeley suggests an expression that might approximate the total branch length for arbitrary n and D:

$$E[T_B(\alpha_1, \alpha_2, \ldots, \alpha_d)] \approx 4\left(\sum_{i=1}^{n-1} \frac{1}{i} + \frac{1}{4R^*} \sum_{i=1}^{d-1} \frac{1}{i}\right), \tag{2.50}$$

where $R^* = NDr/(D-1)$ is the scaled relocation probability of reaching one particular other deme, and d is the the number of demes that the sequences are sampled from. The accuracy of the approximation depends on D, n and how the

lineages are sampled among the demes. For $n = 4$ sampled from one deme, the approximation is fairly accurate, and the error is at most 1%. As the number of demes approaches infinity the value for the sample configurations $(1, 1, \ldots, 1)$ and $(2, 1, \ldots, 1)$ converge to the value obtained by (2.50)

The few existing exact results and approximations are for small samples and very simple model settings. Beyond this level of complexity, and for large samples, simulation is the most appropriate approach. For a model of two demes, Tajima (1989) has simulated the dependency of the mode of sampling for the number of segregating sites. The figures 2.5 through 2.10 summerises some of his results. Below, the main features of the figures are listed. $S(i, j)$ designates the expected number of segregating sites in a sample of i sequences taken from deme one and j sequences taken from deme two. Note that the figures use different notation for population sizes and relocation probabilities.

Isotropic and Conservative Migration: (figure 2.5) The effects of sampling are symmetric because of the complete symmetry of the model. The values for $S(n/2, n/2)$ increase as R decrease, and the effect of coalescence preclusion becomes stronger. $S(n, 0)$ and $S(0, n)$ are smallest for $R = 1$. This is where the effect of early coalescences is largest compared to the effect of coalescence preclusion.

Conservative Migration $N_1 < N_2$: (figure 2.6) As R decreases $S(n, 0)$ decreases since the lineages spend more time in the small deme they are sampled in. As R decreases $S(0, n)$ increases because the lineages spend more time in the large deme. Note in addition, that they do not decrease and increase at the same rate. This is due to the fact that the relocation probabilities are not the same. We have that $r_1 N_1 = r_2 N_2$ implying that $r_1 > r_2$. Hence, if N is constant and R becomes a factor smaller, the r_1 will decrease more in absolute value than r_2

Isotropic Non-Conservative Migration: (figure 2.7) As the relocation rates increase, the values converge to those expected in a panmictic population of size $4N_1N_2/(N_1 + N_2)$. That is, as the the dependence on sampling decreases, the total branch length behaves as in one population with size equal to the harmonic mean of the deme sizes.

Unidirectional Relocation into a larger deme $N_1 < N_2$: (figure 2.8) As R increases, the values of $S(n, 0)$ increases converging to the values of $S(0, n)$ as $R \to \infty$, in this limit case, all lineages will instantaneously relocate to deme two. $S(0, n)$ of cause is unaffected by migration.

Unidirectional Relocation into a smaller deme $N_1 > N_2$: (figure 2.9) As R increases, the values of $S(n, 0)$ decreases converging to the values of $S(0, n)$ as $R \to \infty$. $S(0, n)$ is of cause unaffected by migration.

Unidirectional Relocation into a deme of the same size: (figure 2.5) $S(n, 0)$ decreases as R increases. $S(n, 0)$ is large when R is small. $S(0, n)$ is of cause unaffected by migration.

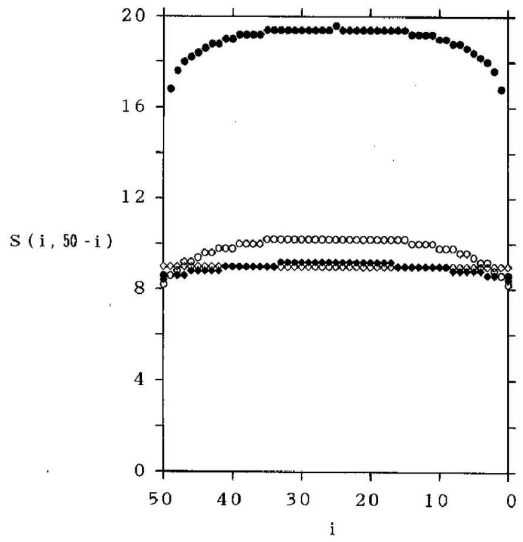


Figure 2.5: Expected number, $S(i, 50-i)$, of segregating sites in a sample of 50 sequences among which i are sampled from deme 1 and $50 - i$ are sampled from deme 2. $R_i = 4N_i r_{ij}$, $\theta_i = 4N_i\mu$. $\theta_1 = \theta_2 = 1$ and $R_1 = R_2$ are assumed. •, $R_1 = 0.1$; ∘, $R_1 = 1$; ♦, $R_1 = 10$; ◇, $R_1 = \infty$.

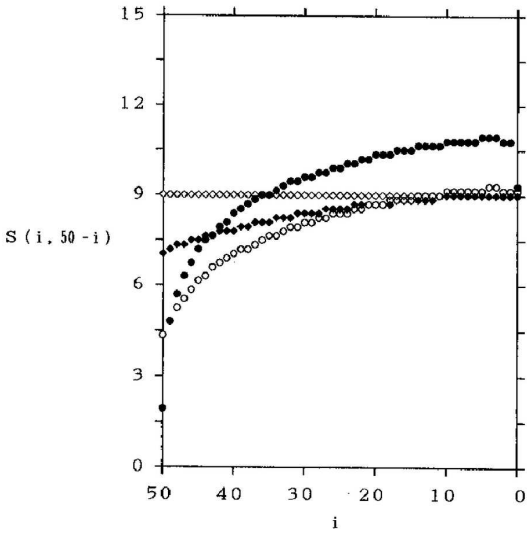


Figure 2.6: Expected number, $S(i, 50-i)$, of segregating sites in a sample of 50 sequences among which i are sampled from deme 1 and $50 - i$ are sampled from deme 2. $R_i = 4N_i r_{ij}$, $\theta_i = 4N_i\mu$. $\theta_1 = 0.1$, $\theta_2 = 1.9$ and $R_1 = R_2$ are assumed. •, $R_1 = 0.1$; ∘, $R_1 = 1$; ♦, $R_1 = 10$; ◇, $R_1 = \infty$;

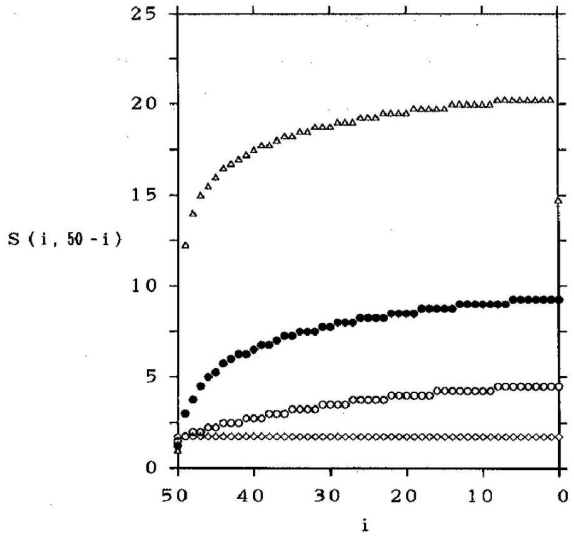


Figure 2.7: Expected number, $S(i, 50-i)$, of segregating sites in a sample of 50 sequences among which i are sampled from deme 1 and $50 - i$ are sampled from deme 2. $R_i = 4N_i r_{ij}$, $\theta_i = 4N_i\mu$. $\theta_1 = 0.1$, $\theta_2 = 1.9$ and $R_2 = 19R_1$ are assumed. $\triangle$, $R_1 = 0.01$; $\bullet$, $R_1 = 0.1$; $\circ$, $R_1 = 1$; $\Diamond$, $R_1 = \infty$;

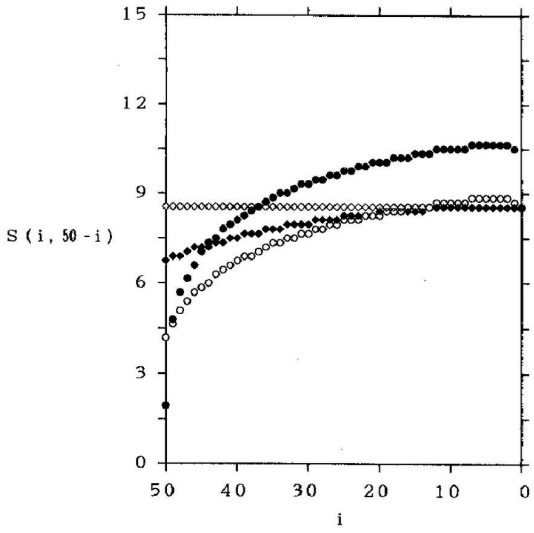


Figure 2.8: Expected number, $S(i, 50-i)$, of segregating sites in a sample of 50 sequences among which i are sampled from deme 1 and $50 - i$ are sampled from deme 2. $R_i = 4N_i r_{ij}$, $\theta_i = 4N_i\mu$. $\theta_1 = 0.1$, $\theta_2 = 1.9$ and $R_2 = 0$ are assumed. $\bullet$, $R_1 = 0.1$; $\circ$, $R_1 = 1$; $\blacklozenge$, $R_1 = 10$; $\Diamond$, $R_1 = \infty$;

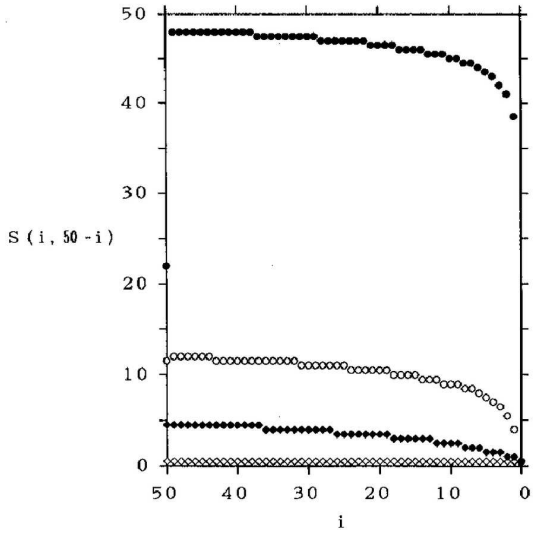


Figure 2.9: Expected number, $S(i, 50-i)$, of segregating sites in a sample of 50 sequences among which i are sampled from deme 1 and $50 - i$ are sampled from deme 2. $R_i = 4N_i r_{ij}$, $\theta_i = 4N_i\mu$. $\theta_1 = 1.9$, $\theta_2 = 0.1$ and $R_2 = 0$ are assumed. •, $R_1 = 0.1$; ∘, $R_1 = 1$; ♦, $R_1 = 10$; ◇, $R_1 = \infty$;

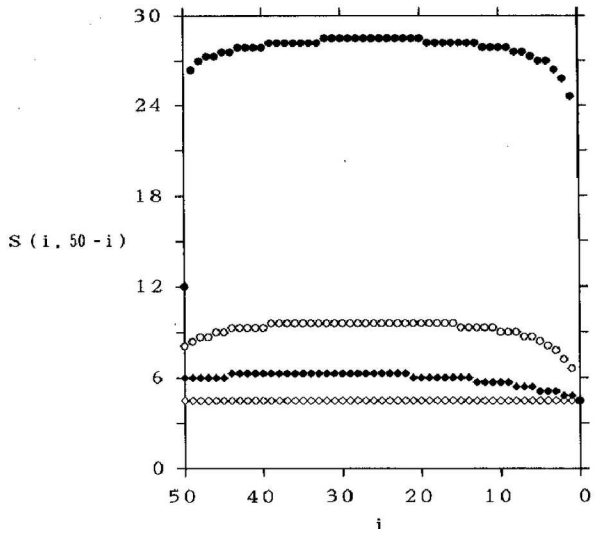


Figure 2.10: Expected number, $S(i, 50-i)$, of segregating sites in a sample of 50 sequences among which i are sampled from deme 1 and $50 - i$ are sampled from deme 2. $R_i = 4N_i r_{ij}$, $\theta_i = 4N_i\mu$. $\theta_1 = \theta_2 = 1$ and $R_2 = 0$ are assumed. •, $R_1 = 0.1$; ∘, $R_1 = 1$; ♦, $R_1 = 10$; ◇, $R_1 = \infty$;

2.4.3 Effect of different deme sizes

The effective sizes of demes determine the coalescence rate in each deme. As noted above, unidirectional relocation into a smaller deme will aggregate lineages under a stronger drift regime, and thus shorten the last branch levels. Analogously unidirectional relocation into a larger deme will prolong the last branch levels (in addition to the effect of coalescence preclusion. If relocation is not unidirectional, the effects of deme sizes become difficult to untangle from effects pertaining to the relocation regime, such as coalescence preclusion, and early coalescences/aggregation of lineages. If relocation probabilities are very large, each deme size can no longer have an separate effect on tree structure, since in this case, the time spent in one deme between relocation events is very small.

2.4.4 The Strong Migration Limit

The limits of many of the results listed above indicate that the effect of structuring declines as migration becomes large. That is, the relocation waiting time becomes smaller. In the limit where migration is infinitely large, the waiting time is infinitely small. This limit, the strong migration limit, was first investigated by Nagylaki (1980), who showed that the ancestral relationship of lineages in the population in this case behaves as in a panmictic population. However, the effective population size, N_e, and thus the expected coalescence time, is smaller than of equal to N_T. Nagylaki designated the resulting effective population size the migration effective population size.

Formally, the strong migration limit is obtained by passing $N_i \to \infty$ for all i without making the assumption that $\lim_{N_i \to \infty} r_{i*} N_i = R_{i*}$. In other words, the backwards migration matrix is held constant as the deme sizes go to infinity. Since this implies that as $R_{i*} \to \infty$, a relocation event is infinitely more probable than a coalescence event. As a result, there are infinitely many relocation events between each coalescence event. This means, that the spacial distribution describing the probability of finding a lineage in the different demes, is stationary. This distribution can also be interpreted as describing the fraction of the time that a lineage will be located in the different demes.

This may not be meaningful in a biological sense, but the limit has properties that can be exploited, if a model can be approximated to it. For the approximation to be justified, each R_{i*} does not have to be large per se, only so much larger compared to the coalescence rate, that we can assume that we have so many relocation events between each coalescence event, that the result is effectively the same.

Migration Effective Population Size

Nordborg (1997) gives a simple and intuitive interpretation of the migration effective population size for two demes and a sample of two sequences. If the time scales of the Coalescent process and the relocation process can be assumed to be separate as described above, the fraction of the time that a lineage is located in

deme one is equal to the normalised rate at which lineages relocate into deme one, $r_{21}/(r_{12}+r_{21})$. Since the location of the lineages are independent, the fraction of the time that both sequences are located in deme one is $r_{21}^2/(r_{12}+r_{21})^2$. The lineages can only coalesce when they are located in the same deme, and when they are, they will do so at a rate $1/c_i$. That means that the coalescence rate can be expressed as the sum of the coalescence rates of the two demes, each weighted by the time, that both lineages are expected to be located together in the deme:

$$\lambda = \frac{r_{21}^2}{(r_{12}+r_{21})^2}\frac{1}{c_1} + \frac{r_{12}^2}{(r_{12}+r_{21})^2}\frac{1}{c_2} = 1 + \frac{(r_{12}c_1 - r_{21}c_2)^2}{(r_{12}+r_{21})^2 c_1 c_2}. \tag{2.51}$$

Since $N_e = N_T/\lambda$, N_e will always be equal to or smaller than one. Clearly, it will be one only if $r_{12}c_1 = r_{21}c_2$. In this case migration is conservative.

Since the Coalescent behaves as in a panmictic population of size N_e in the strong migration limit, the Coalescent is a standard Kingman one, with population size N_e. This implies that the coalescence rate in units of N_e for k lineages, is simply $\binom{k}{2}$

Nagylaki's more general formulation of the migration effective population size is

$$N_e = \frac{1}{\lambda}N_T, \quad \lambda = \sum_{i\in S} \nu_i^2/c_i, \tag{2.52}$$

(Nagylaki 1980), where $c_i = N_i/N_T$ and $\nu = \{\nu_1 \dots \nu_D\}$ is the stationary spacial distribution of a lineage. Hence, ν_i is the probability of finding one of the lineages in deme i. ν and 2 are thus parameters in the a multinomial distribution describing the probability of a particular distribution of the lineages. ν is obtained as the left eigenvector of the backward migration matrix corresponding to the eigenvalue one[1]. To explain, λ is the sum of the probabilities, that two lineages are found in the same deme (the fraction time they spend together in that deme) multiplied by the coalescence intensity for two lineages in that deme. $1/\lambda$ can be expressed as a harmonic mean, and since $\sum_i c_i = 1$ we have that

$$1/\lambda = 1 \Big/ \sum_i \frac{\nu_i}{(c_i/\nu_i)} \leq \sum_{i\in S} \nu_i(c_i/\nu_i) = 1, \tag{2.53}$$

since the harmonic mean is always less than or equal to the arithmetic mean. Hence, $\lambda \geq 1$ and $N_e \leq N_T$ with equality if and only if $\nu = c$, i.e. that we have for all demes, that the probability of finding a lineage in a deme is equal to the fraction of the total population size which that deme constitutes. If $\mathcal{R}$ denote the backward migration matrix, we have that $\nu^T\mathcal{R} = \nu^T$ (ν^T is the pranspose of ν), implying that $\nu_i = \sum_j \nu_j r_{ji} = \nu_i r_{ii} + \sum_{j:\, j\neq i} \nu_j r_{ji}$ Hence, $N_e = N_T$ if and only if

[1] A left eigenvector with corresponding eigenvalue one, is the stationary distribution for a matrix, since that eigenvector can be multiplied by the matrix without changing.

$$c_i r_{ii} + \sum_{j \in S: j \neq i} c_j r_{ji} = c_i. \tag{2.54}$$

Multiplying by N_T on both sides, we see that this is only fulfilled when migration is conservative. i.e. when the sum of all scaled relocation probabilities including the one representing no relocation, is equal to the size of the deme.

With isotropic migration and uneven deme sizes, the migration effective population size is given by

$$N_e = DN^{(h)}, \tag{2.55}$$

where $N^{(h)}$ denotes the harmonic mean of the deme sizes (Nagylaki 1998). This result is effectively the same as that of Slatkin (2.31), just for strong migration.

As far as the robustness of the strong migration approximation is concerned, recall that Takahata showed that with even deme sizes and isotropic migration, the strong migration approximation is valid as long as $4Nr \geq 10$.

Total Tree Length

The expected time, in units of N_T, to the most recent common ancestor of the sample, is not surprisingly given by

$$\lim_{\substack{R_{i*} \to \infty \\ i \in S}} E(T_{n \to 1}) = \left(\sum_{i \in S} \frac{\nu_i^2}{c_i} \right)^{-1} 2 \left(1 - \frac{1}{n} \right), \tag{2.56}$$

(Notohara 1993), and in units of N_e as given by (2.52) the familiar result for the Kingman Coalescent is obtained

$$\lim_{\substack{R_{i*} \to \infty \\ i \in S}} E(T_{n \to 1}) = 2 \left(1 - \frac{1}{n} \right), \tag{2.57}$$

(Kingman 1982*b*).

Location of the Common Ancestor

If the population is subdivided into D demes, the Markov process of the structured Coalescent has D absorbing states, each corresponding to a common ancestor in some of the D demes: $\varepsilon^i = \{\alpha \in I;\ i \in S;\ \alpha_i = 1;\ 0 \text{ otherwise}\}$. Following from the fact that in the strong migration limit, the distribution of the lineages among the demes is stationary, the probability of some particular location of the sample is independent of any previous location. This implies, that the location of the last two lineages is independent of the location of lineages before the point in time where the sample coalesced into these two remaining lineages. As a result, the absorbing state of the Markov process is determined solely by the stationary spacial distribution of the last two lineages, and the probabilities that the lineages

will coalesce when located in the same deme. In conclusion, the probability of an absorbing state ε^i is given by

$$\lim_{\substack{R_{i*}\to\infty \\ i\in S}} \mathbf{P}(\varepsilon^i|\ \alpha) = \frac{(\nu_i^2/c_i)}{\sum_{j\in S}(\nu_j^2/c_j)} \text{ for all } \alpha \in I,\ |\alpha| \geq 2, \tag{2.58}$$

(Notohara 1993).

2.4.5 The Large D Approximation

The large D approximation, (Wakeley 1998, 1999 and 2001), is based on the assumption that the number of demes is so large, that it is very unlikely that a lineage relocates to an occupied deme, and thus that the probability that more than two lineages collect in one deme is negligible.

Wakeley divides the ancestral process into a scattering phase and a collecting phase. The scattering phase is the process ongoing from time zero until the all remaining lineages are in separate demes. During this phase, only the probability of coalescences between lineages in the same demes and relocation events to unoccupied demes, need to be taken into account. Sampling n_i sequences in deme i, the probability of having n_i' lineages in deme i at the end of the scattering phase is

$$\mathbf{P}(n_i'|\ n_i) = \frac{S_{n_i'}^{(n_i)}(2M_i)^{n_i'}}{(2M_i)_{(n_i)}}, \tag{2.59}$$

where $M_i = \sum_j 2Nr_{ij}$, $S_j^{(i)}$ is an unsigned Sterling number of the first kind, and $x_{(r)} = x(x-1)\cdots(x+r-1)$. Since the events in the different demes are independent, the density of the sample at the end of the scattering function is $\mathbf{P}(\mathbf{n}'|\ \mathbf{n}) = \prod_{i=1}^{n} \mathbf{P}(n'|\ n)$

The collecting phase is a Markov process of relocations between unoccupied demes, punctuated by rare relocation events to demes already occupied by a lineage. Analogous to the approach taken by Takahata (1991) for his low migration approximation, Wakeley assumes, the collecting phase is so much longer than the scattering phase, that the entire process can be approximated by a description of the collecting phase. This assumption is valid if the number of demes is large compared to the number of sequences in the sample.

Let p denote the stationary distribution of each lineage among the demes (see section 2.4.4). This is a multinomial distribution with parameters one and $\pi = \{\pi_1 \ldots \pi_D\}$ where π_i is the probability of finding a lineage in deme i. Let r_{i*} denote the probabilities of a relocation event from deme i to any other deme, and e_{*j} the average probability that a relocation from some deme $i \neq j$ will be a relocation to deme j. The rate, ω, of relocations into occupied demes, is then given by

$$\omega = 2\binom{n'}{2}\sum_{i\in S} r_{i*}\,\pi_i \sum_{j\in S} e_{*j}\,\pi_j, \tag{2.60}$$

where n' denotes the total number of demes left after the scattering phase. To explain, this is the average probability that a particular lineage is located in one particular deme and that it relocates from that deme, times the average probability that it relocates to one particular deme and that one particular lineage resides in this deme, times the number of pairs this can happen to, times two because the relocation event may be in both directions. The time until the the next relocation into an occupied deme is exponentially distributed with parameter ω.

The probability that the deme, in which the two demes meet, is a deme of type i is

$$f_i = \frac{e_{*i}}{\sum_{i\in S} e_{*j}}\,\pi_i. \tag{2.61}$$

When this eventually happens, the lineages may coalesce before one of them relocates again. The probability of this outcome, is a simple waiting of the exponential intensities $1/(1+2N_i r_{i*})$. The average probability that coalescence event follows from a relocation to an occupied deme is

$$\sum_{i\in S} \frac{1}{1+2N_i r_{i*}}\, f_i. \tag{2.62}$$

The expected number of these punctuating events that elapse before a coalescence takes place is geometrically distributed with parameter $N^{-1}/(N^{-1}+2r)$.

Multiplying the probability of a relocation into an occupied deme (2.60) by the probability that the outcome of such an event is a coalescence (2.62) the intensity of the exponentially distributed time to a coalescence in the collecting phase becomes

$$\lambda = 2\binom{n'}{2}\sum_{i\in S} r_{i*}\,\pi_i \sum_{j\in S} e_{*j}\,\pi_j \sum_{i\in S} \frac{1}{1+2N_i r_{i*}}\, f_i. \tag{2.63}$$

If N denotes the arithmetic mean of the deme sizes, then measuring time in units of ND and letting $D\to\infty$, a Kingman Coalescent results with effective population size

$$N_e = \left(2\sum_{i\in S} r_{i*}\,\pi_i \sum_{j\in S} e_{*j}\,\pi_j \sum_{i\in S} \frac{1}{1+2N_i r_{i*}}\, f_i\right)^{-1}. \tag{2.64}$$

This means, that in the limit, where the number of demes is much larger than the sample, the last part of the tree will behave as a standard Kingman coalescent for a population size of N_e. This collecting phase will comprise the entire tree if only one sequence is sampled in each sample deme.

To find some specific measure e.g the time to the most recent common ancestor, the time to the second coalescence or the total branch length, the result must be averaged over all possible outcomes of the scattering phase. Hence for some sample configuration, $\mathbf{n}$, the time to the most recent common ancestor is given by

$$\mathbf{P}(T_T(\mathbf{n}) = t) = \sum_{\mathbf{n}'} \mathbf{P}(T_T(\mathbf{n}') = t)\, \mathbf{P}(\mathbf{n}' |\, \mathbf{n}) \tag{2.65}$$

Summing over all possible values of t give the expectation of T_T

$$E[T_T(\mathbf{n})] = \sum_{\mathbf{n}'} E[T_T(\mathbf{n}')]\, \mathbf{P}(\mathbf{n}' |\, \mathbf{n}) \tag{2.66}$$

This approach obviously eases the computational problems otherwise encountered when approaching the structured Coalescent. Wakeley compared the approximation to simulations and concluded that the large D approximation is good as long as the number of demes is at least three times the sample size.

2.4.6 Source-Sink Populations

The case of asymmetric migration is often denoted source-sink migration. However considering a source-sink system only as a system of asymmetric migration with constant deme sizes is insufficient. The underlying reason for the source-sink dynamic must be taken into account as well. This reason, as obvious as it may seem, is that some demes are overproducers compensating for the underproduction in other demes. These local demographic differences will result in local differences in genetic drift, and thus influence the coalescence process. A more precise definition of a source-sink functionality is: Asymmetric migration among demes *resulting* from demographic differences among demes, serving to distribute surplus individuals from overproducing demes between under-producing demes.

Pulliam (1988) drew attention to the fact that for many populations, a large fraction of the individuals may be located in sink demes, and that a small source deme may potentially supply a large collection of sink demes.

A source-sink functionality as defined above, taking the underlying demographic differences among demes into account, has not been modelled in a Coalescent framework. The following chapter will address the problem of source-sink populations to its full extent, and investigate the effects of these causative demographic differences among demes.

Chapter 3

The Coalescent in Source-Sink Populations

This chapter is devoted to investigate asymmetric migration in a system of conserved deme sizes. A system where migration is non-conservative in the sense described above, but where deme sizes are nevertheless in equilibrium. In such a source-sink system, some demes are overproducers and others are under-producers. Migration upholds the dynamic equilibrium of deme sizes by distributing the surplus individuals among sinks. Hence, the equilibrium is a consequence of the demographic properties of the demes.

From a data set we can maximally obtain the backwards migration matrix, i.e. in the composite parameters $N_e r_{ij}$, and the fraction sizes of the demes, c_i. N_e denotes the total effective population size. Hence, the effects of demography on local drift regimes due to a source-sink functionality between demes, can not be distinguished from other effects on genetic drift and thus on effective deme sizes, since the contribution of each effect will be confounded by the composite nature of effective population size.

The Wright-Fisher model describes a population down to the composite parameters listed above that may maximally be obtained from a data set. Here, however, I aim to describe how demographic differences between demes may affect effective deme sizes and effective population size. Since these effects can not be separated from other effects on effective deme sizes, the scope of the Wright-Fisher model is not sufficient to investigate this.

A source-sink dynamic must be considered in a Moran model, since this incorporates the demographic parameters in question. Below a model with a scope adequate to describe the nature of a source-sink functionality will be presented. In brief, this is done by formulating a model that enables a separate investigation of the effects of migration rates, deme sizes and demography effects.

The effect of varying demographic parameters among demes, on the divergence of sampled sequences, will be considered in detail.

3.1 Setting the Scene

I adopt the Moran model, that unlike the Fisher-Wright incorporates the birth and death parameters, so crucial for this kind of modelling. Specifically we will consider the situation where a set of two demes $S = \{1, 2\}$ each of constant size N_i exchange migrants. The growth rates differ between demes, but asymmetric migration distributes surplus individuals among sinks, thereby conserving deme sizes.

In the Moran model, every unit time one lineage is randomly sampled to die and one is randomly sampled to give rise to a new lineage. Hence, each time unit, one lineage is copied and takes the place of one that dies. It is assumed, that the lineage that dies may leave an offspring, and that an offspring may take the place of the mother lineage. This implies that it does not matter whether the lineage to die or the lineage to be split is sampled first. The birth and the death event may each occur in any of the D demes.

Right after the birth and the death event a migration event may occur. The coupling between the birth-death process, and the migration process determines the stochasticity with respect to the equilibrium of deme sizes. In the model presented here, these processes are completely coupled. That is, if the lineage that dies and the one that is split are located in different demes, the surplus individual migrates to the deme that is short of one individual. This precludes any fluctuation of deme sizes. Migration is stochastic, as we shall see shortly, since this is a function of the stochastic birth and death parameters.

Every deme is associated with three parameters: A per capita birth parameter, β, a per capita death parameter, δ, and a deme size, N. Hence, β is the number of births per time per lineage, and δ is the number of deaths per unit time per lineage. Thus $\beta - \delta$ is the growth rate of the deme. If β or δ are fraction numbers, it is assumed that the individuals are added or removed randomly one at a time, so that the parameters averagely hold. The deterministic parameters are kept for convenience.

The expected life time of an individual in the Moran model (with continuous time) is the mean of an exponential distribution. This implies that for any age, there is a positive probability that a chosen individual becomes older. Whether this is an appropriate description of life expectancy is arguable. If, however, it is assumed that ageing plays a minor role relative to death by competition, predation or disease, an exponential description seems reasonable.

Given the wider scope of the Moran model compared to the Wright-Fisher model, structure can be modelled as a function of of birth and death rates in demes and the covariance of birth and death events in different demes, and not only as a function of migration as in the Wright-Fisher model.

3.2 Formulation of a Structured Moran Model

When the newborn individual to be added to a deme in a Moran event is sampled it is done from the collection of demes, weighted with the birth parameters and the size of each deme, so that the probability of a birth event in deme i is:

$$P(B_i) = \frac{\beta_i N_i}{\sum_{k \in S} \beta_k N_k} = b_i, \tag{3.1}$$

implying that

$$\sum_{i \in S} b_i = 1. \tag{3.2}$$

Analogously the probability of a death event in deme i is:

$$P(D_i) = \frac{\delta_i N_i}{\sum_{k \in S} \delta_k N_k} = d_i, \tag{3.3}$$

so that

$$\sum_{i \in S} d_i = 1. \tag{3.4}$$

In the following it will be assumed that the value of *either* b *or* d are constant over all demes. This is to ensure, that all demes, on average, are involved in events equally often. Since time is measured in terms of events, this assumption assures that time has the same meaning in all demes. The model is not confined to this premise. It is only posited to keep the results as simple as possible. This imposes some limitations on the composite parameter $\beta_i N_i$, that must be kept in mind. (3.1) and (3.3) will be denoted the birth and the death rate respectively. The one of them that is allowed to vary among demes will be referred to as the primary parameter and the one that is equal over all demes is denoted the secondary parameter. The model where the death rates are even, is denoted the death model, and the model where the birth rates are equal is denoted the birth model. Note that in these models the per capita parameter is only even among demes if all deme sizes are equal. In the formulation of the model deme sizes are allowed to vary. However, for the sake of simplicity, explanation of results will assume equal deme sizes. The effects of different sized deme sizes will be considered in a separate section.

3.2.1 Sampling in an Unstructured Setting

In the unstructured situation there is independence of where the birth event and the death event takes place. Hence, the probability that one lineage is split into two that stay in deme one is a simple product of the probabilities of a death event and a birth event in deme one:

$$P(B_1, D_1) = b_1 d_1. \tag{3.5}$$

Analogously the probability that in a Moran event we have a migration from deme one to deme two is

$$P(B_1, D_2) = b_1 d_2, \tag{3.6}$$

and the probability that we have a migration from deme two to deme one is:

$$P(B_2, D_1) = b_2 d_1. \tag{3.7}$$

Hence, the simultaneous density of B and D is

D_2	$b_1 d_2$	$b_2 d_2$
D_1	$b_1 d_1$	$b_2 d_1$
	B_1	B_2

(3.8)

3.2.2 Sampling in a Structured Setting

In a model with structure the events of birth and death are no longer independent. The more pronounced the structure, the stronger the dependence. Let deme one be the *sink* and deme two the *source*. The density of the vector (B, D) in the birth model and in the death model can be expressed as

D_2	$b_1(1-s)$	$b_2 p$		and	D_2	$d_2(1-g)$	$d_2 g$
D_1	$b_1 s$	$b_2(1-p)$			D_1	$d_1 h$	$d_1(1-h)$
	B_1	B_2				B_1	B_2

(3.9)

respectively. In the birth model the d's are substituted for terms of s and p. This way the birth and death rates can be varied. The values of s and p determines the extent to which the sampling of lineages to die are skewed away from the unstructured panmictic situation. In the maximally structured situation, there is only migration from the source to the sink, implying that s is one. In this case p must equal d_2/b_2 if the double stochasticity is to be retained. In the unstructured case s equals d_1, and p equals d_2. The limits to g and h in the death model are determined analogously. To summerise $d_1 \leq s \leq 1,\ d_2 \leq p \leq d_2/b_2,\ b_1 \leq h \leq b_1/d_1$ and $b_2 \leq g \leq 1$. The covariance of B and D under the birth and the death model are

$$\text{Cov}(B, D) = b_1 b_2 (p + s - 1) = b_1 (s - d_1) = b_2 (p - d_2) \tag{3.10}$$

and

$$\text{Cov}(B, D) = d_1 d_2 (h + g - 1) = d_2 (g - b_2) = d_1 (h - b_1) \tag{3.11}$$

respectively. The last two terms in (3.10) and (3.11) are obtained by expressing the densities (3.9) only in term of one structure parameter.

To get the same degree of source-sinkness in the birth and the death model the value of b_1 under the death model must equal the value of d_2 under the birth model, and b_2 under the death model must equal the value of d_1 under the birth model. Since we have that under the birth model $b_2 p + b_1(1 - s) = d_2$ and under the death model that $d_1 h + d_2(1 - g) = b_1$ we have that $p + s = h + g$. Hence on the condition that both models describe the same over/underproduction situations in the demes, (3.10) equals (3.11), and the two models collapse into one. This means that the density of (B,D) for both models can be expressed as the doubly stochastic matrix:

$$\begin{array}{c|cc} D_2 & b_1 d_2 - Cov(B,D) & b_2 d_2 + Cov(B,D) \\ D_1 & b_1 d_1 + Cov(B,D) & b_2 d_1 - Cov(B,D) \\ \hline & B_1 & B_2 \end{array} \tag{3.12}$$

In the case of maximal covariance, we have a maximally structured scenario. Here the covariance has the maximal value of $b_1 d_2$ and there is only migration from the source to the sink. That is, the minimal migration needed to compensate for the differences in demography between the two demes. In this limit the density of the vector (B, D) is:

$$\begin{array}{c|cc} D_2 & 0 & d_2 \\ D_1 & b_1 & b_2 d_1 - b_1 d_2 \\ \hline & B_1 & B_2 \end{array} \tag{3.13}$$

A deviation from the maximal covariance will result in additional symmetric migration between the demes. Hence, the migration between the demes can be separated into an unidirectional compensating migration serving to conserve deme sizes and a symmetric mixing migration serving only to mix the lineages in the demes.

In the maximally structured situation the probability of a split in the source is governed solely by d_2 in the source, and by b_1 in the sink, since $\mathbf{P}(B_2|D_2) = 1$ and $\mathbf{P}(D_1|B_1) = 1$. In other words, maximal structure corresponds to a situation where, if possible, an empty space in a deme will always be filled with an individual born in that same deme. Hence, maximal structure is the situation where we have as many split events with no migration as the the smallest of the parameters allow.

The diagonal entries in (3.12) represent both the probability of an immigration event and an emigration event, depending on which deme is considered. Here however, since these are equivalent descriptions in a two-deme system, all migration rates can be expressed in terms of immigration rates. The matrix describing the probabilities of migration events is:

$$F = \begin{bmatrix} 1 - b_i d_j - Cov(B,D) & b_i d_j - Cov(B,D) \\ b_j d_i - Cov(B,D) & 1 - b_j d_i - Cov(B,D) \end{bmatrix} \tag{3.14}$$

Note that the probabilities in this matrix are the probabilities of migration of some lineage in a particular deme, and not the probabilities om migration of some particular lineage.

This way of describing migration allows us to separate the part of the migration rate that is needed to conserve deme sizes, and the part that is symmetric, serving only to mix the lineages between demes. The compensating migration is the unidirectional migration in the maximally structured scenario, where the covariance has the maximal value. The mixing migration is given by the departure form the maximal covariance. Thus the relative size of the two decides to what extent a deme is an over or under producer, and to what extent it is a high or low turnover deme.

3.2.3 The Coalescent in Two Demes

Under the retrospective Coalescent model we must consider the backwards transition probabilities. Since the matrix (3.12) is double stochastic, the probability that a particular lineage in a deme is split into two in the forward model, equals the probability, that backwards in time, a pair of lineages in the deme coalesce into one.

In the Moran model, for each migration event in the forward model, both a donor and a receptor deme for the migration is given by where the birth and the death event is sampled. Hence, owing to the double stochasticity of the density (3.12), a the probability of a forward migration event from deme j to deme i equals the probability of a backwards relocation event from deme i to deme j. The matrix describing the probabilities that some lineage in a deme relocate backwards in time is thus:

$$H = \begin{bmatrix} 1 - b_j d_i - Cov(B,D) & b_j d_i - Cov(B,D) \\ b_i d_j - Cov(B,D) & 1 - b_i d_j - Cov(B,D) \end{bmatrix} \tag{3.15}$$

It must again be stressed that these are the probabilities, that some lineage in a particular deme relocates, and not the probability of the relocation of a particular lineage. Below the transition probabilities pertaining to the particular lineages from the sample are considered.

For two demes and allowing either b or d the probability of a coalescence event between two lineages in deme i is:

$$\frac{2\binom{\alpha_i}{2}(b_i d_i + Cov(B,D))\varphi}{{N_i}^2} \quad i \in \{1,2\}, \tag{3.16}$$

where α_i is the number of lineages from the sample in deme i. To explain, this is the probability, that two lineages in the deme coalesces, times the probability, of drawing two particular lineages from the deme , $2/{N_T}^2$, times the number of pairs in our sample, $\binom{\alpha_i}{2}$. The correction factor $\varphi = N_i/(N_i - 1)$ designates the probability that it is *not* the same lineage that is sampled twice. If the same lineage is sampled twice, it corresponds to the event where a newborn takes the place of the mother, in which case all the lineages are still represented in the deme. In other words, if this is the case, the equivalence classes representing the sample back in time would not be altered. Since $\lim_{N\to\infty} \varphi = 1$, φ can of cause be left out if deme sizes are large. It is included here because large demes are not a prerequisite in the Moran model.

The probability of a coalescence event of a lineage located in deme i to one located in deme j is:

$$\frac{\alpha_i(d_i b_j - Cov(B,D))}{N_i}\left(\frac{\alpha_j}{N_j}\right) \quad i,j \in \{1,2\}. \tag{3.17}$$

That is, the probability that a lineage relocates from deme i to deme j, times the probability the birth event in deme j is a reproduction of a lineage from the sample.

The probability of a relocation from deme i to deme j without affecting any other lineages from the sample is:

$$\frac{\alpha_i(d_i b_j - Cov(B,D))}{N_i}\left(1 - \frac{\alpha_j}{N_j}\right) \quad i,j \in \{1,2\}. \tag{3.18}$$

Time and thus the transition probabilities are scaled with N_T/σ^2. σ^2 is the variance of the joint distribution of $\nu_i \in \{\nu_i, \ldots, \nu_n\}$ where ν_i is the number of offspring of a particular lineage in one event. In the standard Moran Model with no structure, $\sigma^2 = 2/N$, so we scale with $N^2/2$. This is done to make the model congruent with the standard results for the Coalescent, see section 1.2. Hence, (3.16), (3.18) and (3.18) turns into (3.19), (3.21) and (3.21)

$$\frac{\binom{\alpha_i}{2}(b_i d_i + Cov(B,D))\varphi}{{c_i}^2} \quad i \in \{1,2\} \tag{3.19}$$

$$\frac{\alpha_i\alpha_j(d_ib_j - Cov(B,D))}{2c_ic_j} \quad i,j \in \{1,2\} \tag{3.20}$$

$$\frac{\alpha_i(d_ib_j - Cov(B,D))(c_jN_T - \alpha_j)}{2c_ic_j} \quad i,j \in \{1,2\}, \tag{3.21}$$

where c_i is the fraction of the total population size that the deme constitutes. (3.19) will be referred to as the in-deme coalescence rate in deme i, (3.20) as the cross coalescence rate from deme i to deme j, and (3.21) as the relocation probability from i to j. The matrix of relocation probabilities, will be referred to as the backwards migration matrix.

The sequence of events up to a coalescence event is a Markov process since the transition probabilities are only dependent on the state in which the sample is presently in. After a coalescence event the number of possible states decreases, and a new Markov process takes the process ahead. We have $l+3$ states, where l is the number of lineages left in the sample. The first $l+1$ states, indexed by α_1 designating the number of the l lineages present in deme one. So $l - \alpha_1 = \alpha_2$. There are further, two states, $l+2$ and $l+3$, that each represent a coalescent in deme one or deme two respectively. Only the process of decreasing the number of ancestors is considered. The rows 1 through $l+1$ in the transition matrix $Q = \{q_{ij}\}$ are zero except:

$$q_{\alpha_1,\alpha_1+1} = \frac{\alpha_2(d_2b_1 - Cov(B,D))(c_1N_T - \alpha_1)}{2c_1c_2} \tag{3.22}$$

$$q_{\alpha_1,\alpha_1-1} = \frac{\alpha_1(d_1b_2 - Cov(B,D))(c_2N_T - \alpha_2)}{2c_1c_2} \tag{3.23}$$

$$q_{\alpha_1,l+2} = \frac{\binom{\alpha_1}{2}(b_1d_1 + Cov(B,D))\varphi}{{c_1}^2} + \frac{\alpha_1\alpha_2(d_1b_2 - Cov(B,D))}{2c_1c_2} \tag{3.24}$$

$$q_{\alpha_1,l+3} = \frac{\binom{\alpha_2}{2}(b_2d_2 + Cov(B,D))\varphi}{{c_2}^2} + \frac{\alpha_1\alpha_2(d_2b_1 - Cov(B,D))}{2c_1c_2} \tag{3.25}$$

$$q_{\alpha_1,\alpha_1} = -(q_{\alpha_1,\alpha_1+1} + q_{\alpha_1,\alpha_1-1} + q_{\alpha_1,l+2} + q_{\alpha_1,l+3}). \tag{3.26}$$

The rows $l+2$ and $l+3$ have all zero entries except $q_{l+2,l+2} = 1$ and $q_{l+3,l+3} = 1$, since these states are absorbing. Hence we have a matrix:

$$Q = \begin{bmatrix} B & C \\ 0 & I \end{bmatrix} \tag{3.27}$$

where the off diagonal entries of B matrix are the relocation probabilities, and the diagonal entries are given by (3.26). C is the matrix of coalescence probabilities, and I is a 2×2 identity matrix. The transition probabilities are all independent of the time elapsed between coalescence events, so the Markov transition matrix is a stationary one. In conclusion the probability of a transition from state i to j is given by

$$p_{ij} = \delta_{ij} + q_{ij}\frac{2}{N_T^2}, \tag{3.28}$$

where δ_{ij} is the Kronecker delta[1]. Let $\mathcal{P} = \{p_{ij}\}$. Since time is scaled in units of $N_T^2/2$, passing all deme sizes to infinity produces the continuous Markov process

$$\lim_{\substack{N_i \to \infty \\ i \in S}} \mathcal{P}^{[\frac{N_T^2}{2}t]} = e^{Qt}, \tag{3.29}$$

where $[\frac{N_T^2}{2}t]$ indicates that time is measured in units of $N_T^2/2$.

In the Moran model per definition only one event can happen in each time unit. This may be a coalescence event or one relocation event. Hence, passing the deme sizes to infinity only serves to change the discrete Markov chain into a continuous process. (In the Wright-Fisher model the diffusion approximation also serves to make the probability of multiple coalescence or relocation events negligible.) In other words the continuous Markov process describing the structured Coalescent, $\mathcal{P}$, is exact in the Moran model. The only approximation involved is the approximation to continuity.

The distribution of time to the first event of any kind is exponential with rate parameter equal to that of the diagonal entries in Q that represents the present distribution, α, of the lineages among the two demes.

$$P(T > t \mid \alpha) = e^{-(q_{i,i})t} \quad \Rightarrow \quad P(T < t \mid \alpha) = 1 - e^{-(q_{i,i})t}. \tag{3.30}$$

In other words, the rate parameter of the exponential distribution is the sum of all the possible transition probabilities given a particular distribution of the lineages:

$$\begin{aligned} P(T < t \mid \alpha) &= 1 - \exp\left(- \left(\sum_{i \in S} \frac{\binom{\alpha_i}{2}(b_i d_i + Cov(B,D))\varphi}{{c_i}^2} \right.\right. \\ &+ \sum_{i \in S} \sum_{j \in S: j \neq i} \frac{\alpha_i \alpha_j (d_i b_j - Cov(B,D))}{2 c_i c_j} \\ &+ \left.\left. \sum_{i \in S} \sum_{j \in S: j \neq i} \frac{\alpha_i (d_i b_j - Cov(B,D))(c_j N_T - \alpha_j)}{2 c_i c_j} \right) t \right) \end{aligned} \tag{3.31}$$

[1]The Kronecker delta, δ_{ij}, equals one if $i = j$ and equals zero otherwise

The mean of an exponential distribution $1 - e^{-\lambda t}$ is $1/\lambda$, so the expected waiting time to a transition to a particular state is the inverse of its transition rate.

Since the time to the first event of either coalescence or relocation is exponentially distributed, the probabilities of the different events, when an event finally occurs, is a simple weighting of probabilities. Formally, if the process is in state i, the probability that the next transition is a transition to state j is

$$\mathbf{P}(\text{Next transition = } i \to j) = \frac{q_{ij}}{\sum_{j=1, j\neq i} q_{ij}}. \tag{3.32}$$

Hence, if $C_{ii\to i}$ is the event of two lineages from deme i coalescing into one in deme i, $C_{ji\to i}$ is the event of one cross coalescence of a lineage in deme j and one in deme i coalescing into one in deme i, and $R_{i\to j}$ is the event of a relocation of a lineage from deme i to j, then the probabilities of the different events are:

$$P(C_{ii\to i}) = \frac{\binom{\alpha_i}{2}(b_i d_i + Cov(B,D))\varphi}{{c_i}^2 \xi} \tag{3.33}$$

$$P(C_{ij\to j}) = \frac{\alpha_i \alpha_j (d_i b_j - Cov(B,D))}{2 c_i c_j \xi} \tag{3.34}$$

$$P(R_{i\to j}) = \frac{\alpha_i (d_i b_j - Cov(B,D))(c_j N_T - \alpha_j)}{2 c_i c_j \xi}, \tag{3.35}$$

where

$$\begin{aligned} \xi &= \sum_{i\in S} \frac{\binom{\alpha_i}{2}(b_i d_i + Cov(B,D))\varphi}{{c_i}^2} \\ &+ \sum_{i\in S} \sum_{j\in S: j\neq i} \frac{\alpha_i \alpha_j (d_i b_j - Cov(B,D))}{2 c_i c_j} \\ &+ \sum_{i\in S} \sum_{j\in S: j\neq i} \frac{\alpha_i (d_i b_j - Cov(B,D))(c_j N_T - \alpha_j)}{2 c_i c_j} \end{aligned} \tag{3.36}$$

For panmixia and equal deme sizes, that is, if the covariance is zero and

$$\frac{b_1}{b_2} = \frac{d_1}{d_2} = \frac{c_1}{c_2} = 1 \tag{3.37}$$

we have a standard Coalescent in one demographicly homogeneous population, since with this assumption we have (assuming infinitely large demes)

$$\xi = \sum_{i\in S} \frac{\binom{\alpha_i}{2} b_i d_i}{{c_i}^2} + \sum_{i\in S} \sum_{j\in S: j\neq i} \frac{\alpha_i \alpha_j d_i b_j}{2 c_i c_j} = \binom{|\alpha|}{2}, \tag{3.38}$$

which is the coalescence rate for $|\alpha|$ sequences in one panmictic population (Kingman 1982*a*). This is the case because sampling of the death and the birth event is independent and occurs with equal probability in all demes, so that the probability of cross coalescence is the same as in-deme coalescence. In a model with dependent sampling of the death and the birth event, ie. a structured model, the scaled relocation probability has to be infinite for the model to behave panmicticly, as we shall see in section 3.4

3.2.4 Coalescence Intensity and Demography

The coalescence rates depend on the demographic parameters b, d and the $Cov(B, D)$. The dependence on $Cov(B, D)$ is straightforward, and is the same irrespectively of whether the primary parameter is b or d: The larger the $Cov(B, D)$, the larger the in-deme coalescence rate, the smaller the relocation probability, and thus the smaller the cross coalescence rate.

The in-deme coalescence rate, however, is in addition dependent on which parameter that is the primary one. In the unstructured setting, the in-deme coalescence rate is lower in the sink and higher in the source if b is the primary parameter. If the primary parameter is d, the in-deme coalescence rates are affected reciprocally. For the maximally structured situation, the probability of a in-deme coalescence is governed solely by d in the source and by b in the sink (see (3.13)).

3.2.5 Demography, Relocation and Deme Size

The asymmetry of relocation probabilities in a source-sink model is a result of the relative sizes of the net over- or under-production in each deme. The net production is $(\beta_i - \delta_i)N_i$. In the death model it is, $\beta_i N_i$ since the per capita death parameter is one in both demes. Hence, the relative sizes of production $b_i = \beta_i N_i / \sum_{k \in S} \beta_k N_k$ is a function of both β an N. This implies that the migration regime in a source-sink population is governed both by deme sizes as well as per capita parameters. This may seem obvious, but it is important to keep in mind in the following.

The relocation probability of a particular lineage from the sample, which is our concern here, is in addition dependent on the size of the deme that the lineage resides in before the relocation. Hence, in considering the relocation probabilities of a particular lineage, it is crucial to distinguish clearly between the demographic contribution given by b, d and $Cov(B, D)$ in (3.15), and the contribution of one particular deme size. The dependence of both is seen in (3.18). The interplay of per capita parameters and deme sizes is best described by a few examples:

A small source with a large per capita overproduction and a large sink with small per capita underproduction, (Recall that the b's and d's each sum to one, which imposes some restrictions on β_i, δ_i and N_i) will result in an asymmetric flux of individuals given by (3.15). The asymmetry in relocation probabilities, however, will be even more asymmetric, than would be expected from the source-sink relationship between the demes. This owes to the dependence of each relocation

probability on one particular deme size. For a given backwards out-flux of lineages a small deme size will give a larger relocation probability. The drift regimes in each deme is obviously dependent on the deme size. Besides this effect, there is the effects of the local drift differences resulting from demography. Hence, depending on whether it is the birth or the death rate that vary between demes, the effect of different deme sizes may either enhance or, to some extent, cancel out with the effect of demography. In an unstructured setting, the drift regimes will be the same between demes if the relation between the deme sizes is equal to the square of the relation between the two values of the primary parameter.

With a large source with a small per capita overproduction and a small sink with large per capita underproduction (the opposite of the case above), will still result in an asymmetric flux of individuals, but this may not be reflected in asymmetry of the relocation probabilities. This again owes to the dependence of each relocation probability on one particular deme size. The relocation probabilities may in fact be symmetric if the asymmetry in deme sizes cancel out with the asymmetry in net flux of individuals given by (3.15). In an unstructured setting, the relocation probabilities will be symmetric if the relation between the deme sizes is the same as the relation between the two values of the primary parameter. Hence even though there is a true source-sink functionality between the demes, this will not show in the backwards migration matrix. Just as explained above, the drift regimes in the demes may be the same if the effect of deme sizes chancel out with the effect of demography. This may be possible in the death model. In the birth model, however, the different deme sizes will accentuate the local drift differences. Hence, we may have source-sink functionality with large differences in drift strength between demes, that does result in asymmetric relocation probabilities.

Unequal deme sizes may result in asymmetric relocation probabilities even though the net flux of individuals is symmetric. In this case the drift regimes in each deme are of cause solely governed by the deme sizes.

In conclusion, since relocation probabilities are composite parameters, it is not possible to distinguish the effects of deme sizes and the effects from the demographical relation between demes.

Since, the primary subject of this study is the effect of demography an not of deme sizes, the graphical representations and explanations in the following will assume equal deme sizes. This is to present the clearest possible picture of the effects of local demographic differences. Recall that this implies that the per capita death rate is equal among demes in the death model, and that the per capita birth rate is equal among demes in the birth model. Hence, in this situation the representation of demography in the relocation probabilities is not obscured by the effects of different deme sizes.

3.3 Coalescence Time for Two Sequences

In this section it will be investigated how the expected coalescence time of two sequences, for different modes of sampling, is affected by local demographic differences. Exact results for expected coalescence times in the structured Coalescent can be obtained recursively. Following Notohara (1990) and Wakeley (1998), stating the process as a Markov chain with X states and infinitesimal generator Q, we have, as described in the previous chapter:

$$E[T_i] = \frac{1}{q_{i*}} + \sum_{j=1, j\neq i}^{X} \frac{q_{ij}}{q_{i*}} E[T_j]. \tag{3.39}$$

In this model all demes are potentially different, with respect to both relocation and coalescence probability. Hence, we can not derive a general exact result for an arbitrary number of demes. However, with this approach we can derive exact results for sets of demes with specified parameter values.

The amount of calculations increase rapidly with the number of demes and sampled sequences, and the results quickly become to complex for any intuitive theoretical value. Below we consider the exact result for two lineages and two demes. Even for this simple scenario, the analytical results are two complex for any explicatory value without the aid of graphical representation.

$$\lambda_i^{(s)} = \frac{(d_i b_i + Cov(B, D))\varphi}{{c_i}^2} \tag{3.40}$$

$$\lambda_i^{(d)} = \frac{(d_i b_j - Cov(B, D))}{2 c_i c_j} \tag{3.41}$$

$$R_{ij}^{(s)} = \frac{(d_i b_j - Cov(B, D))N_T}{2 c_i} \tag{3.42}$$

$$R_{ij}^{(d)} = \frac{(d_i b_j - Cov(B, D))(c_j N_T - 1)}{2 c_i c_j}. \tag{3.43}$$

The superscripts (s) and (d) signifies same and different, and states whether the two lineages are in the same deme or in different demes. Thus $R_{ij}^{(s)}$ is the probability that a lineage in i was resident in j before the previous event, when both lineages are in deme i. $R_{ij}^{(d)}$ is the same just for the situation where one lineage is in deme i and one is in deme j. They differ because $R_{ij}^{(s)}$ does not take a possible cross coalescence into account, whereas $R_{ij}^{(d)}$ does. Analogously $\lambda_i^{(s)}$ is the in-deme coalescence probability of two lineages in deme i, and $\lambda_i^{(d)}$ is the cross coalescence probability of one lineage in deme i to one in deme j. For large population sizes we may assume that $R_{ij}^{(s)} = R_{ij}^{(d)}$. This may ease calculation of more complicated results.

With the notation (2,0) corresponding to sampling two sequences in deme one, the sink, and none in deme two, the source, we have for two sequences:

$$E[T(2,0)] = \frac{1}{2R_{12}^{(s)} + \lambda_1^{(s)}} + \frac{2R_{12}^{(s)}}{2R_{12}^{(s)} + \lambda_1^{(s)}} E[T(1,1)] \tag{3.44}$$

$$E[T(0,2)] = \frac{1}{2R_{21}^{(s)} + \lambda_2^{(s)}} + \frac{2R_{21}^{(s)}}{2R_{21}^{(s)} + \lambda_2^{(s)}} E[T(1,1)] \tag{3.45}$$

$$\begin{aligned} E[T(1,1)] = & \frac{1}{R_{12}^{(d)} + R_{21}^{(d)} + \lambda_1^{(d)} + \lambda_2^{(d)}} \\ & + \frac{R_{12}^{(d)}}{R_{12}^{(d)} + R_{21}^{(d)} + \lambda_1^{(d)} + \lambda_2^{(d)}} E[T(0,2)] \\ & + \frac{R_{21}^{(d)}}{R_{12}^{(d)} + R_{21}^{(d)} + \lambda_1^{(d)} + \lambda_2^{(d)}} E[T(2,0)]. \end{aligned} \tag{3.46}$$

Solving this system of linear equations we get:

$$E[T(1,1)] = \frac{1 + \frac{R_{12}^{(d)}}{2R_{21}^{(s)} + \lambda_2^{(s)}} + \frac{R_{21}^{(d)}}{2R_{12}^{(s)} + \lambda_1^{(s)}}}{\frac{2R_{12}^{(s)}\lambda_1^{(d)} + \lambda_1^{(s)}\lambda_1^{(d)} + R_{21}^{(d)}\lambda_1^{(s)}}{2R_{12}^{(s)} + \lambda_1^{(s)}} + \frac{2R_{21}^{(s)}\lambda_2^{(d)} + \lambda_2^{(s)}\lambda_2^{(d)} + R_{12}^{(d)}\lambda_2^{(s)}}{2R_{21}^{(s)} + \lambda_2^{(s)}}} \tag{3.47}$$

$$E[T(2,0)] = \frac{1 + \frac{2R_{12}^{(s)}\left(1 + \frac{R_{12}^{(d)}}{2R_{21}^{(s)} + \lambda_2^{(s)}} + \frac{R_{21}^{(d)}}{2R_{12}^{(s)} + \lambda_1^{(s)}}\right)}{\frac{2R_{12}^{(s)}\lambda_1^{(d)} + \lambda_1^{(s)}\lambda_1^{(d)} + R_{21}^{(d)}\lambda_1^{(s)}}{2R_{12}^{(s)} + \lambda_1^{(s)}} + \frac{2R_{21}^{(s)}\lambda_2^{(d)} + \lambda_2^{(s)}\lambda_2^{(d)} + R_{12}^{(d)}\lambda_2^{(s)}}{2R_{21}^{(s)} + \lambda_2^{(s)}}}}{2R_{12}^{(s)} + \lambda_1^{(s)}}. \tag{3.48}$$

$$E[T(0,2)] = \frac{1 + \frac{2R_{21}^{(s)}\left(1 + \frac{R_{12}^{(d)}}{2R_{21}^{(s)} + \lambda_2^{(s)}} + \frac{R_{21}^{(d)}}{2R_{12}^{(s)} + \lambda_1^{(s)}}\right)}{\frac{2R_{12}^{(s)}\lambda_1^{(d)} + \lambda_1^{(s)}\lambda_1^{(d)} + R_{21}^{(d)}\lambda_1^{(s)}}{2R_{12}^{(s)} + \lambda_1^{(s)}} + \frac{2R_{21}^{(s)}\lambda_2^{(d)} + \lambda_2^{(s)}\lambda_2^{(d)} + R_{12}^{(d)}\lambda_2^{(s)}}{2R_{21}^{(s)} + \lambda_2^{(s)}}}}{2R_{21}^{(s)} + \lambda_1^{(s)}} \tag{3.49}$$

The striped area in figure 3.1 shows the range of parameter values allowed by the assumptions in the model. Hence, this range of parameter values compose the domain on which the equations (3.47), (3.48) and (3.49) are defined. We have as initial condition, in terms of the forward mode, that some minimum amount of one-way relocation from sink to source is needed to compensate for the differences in growth rates among demes. The larger the differences in the values of the primary

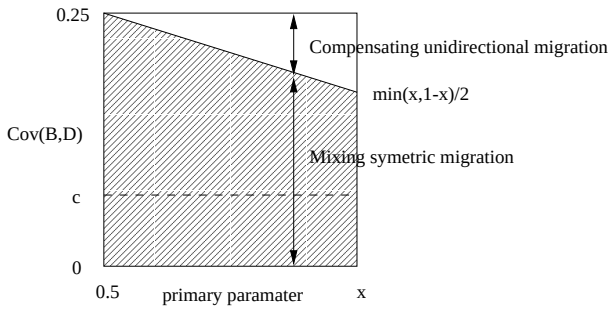


Figure 3.1: The striped area represents the area where the functions in figures 3.2 and 3.3 are defined. This domain is dictated by the assumptions in the model, that some minimum amount of relocation is needed to compensate for the demographical differences. Hence, given the values of b's and d's the $Cov(B, D)$ can only take some maximal value. For a specified set of b and d values these areas also represents the relative sizes of the compensating and mixing part of the relocation probability from sink to source. See the text for an explanation.

parameter among the demes, the larger this compensating relocation has to be. Hence, the values of the primary parameter impose a limit to the maximum level of structuring (maximal covariance), if deme sizes are to be conserved.

As explained in section 3.2.2, the probability of a forward migration event to the sink of any lineage in the source, can be divided into a probability corresponding to the minimal unidirectional migration needed to conserve deme sizes, and a probability corresponding to the additional symmetric migration serving only to mix lineages between two demes. This is also true for the backwards process, and assuming that deme sizes are equal, the relation between the compensating and the mixing migration, is also directly reflected in the relocation probabilities pertaining to the lineages from our sample.

Figure 3.1 shows the relation between these two types of relocation. Consider some value of the primary parameter on the x-axis. For an unstructured scenario, the fraction of the parameter space outside the domain (the white area) corresponds to the fraction of the relocation from sink to source that is compensating, whereas the fraction of the space inside the striped domain corresponds to the fraction of the relocation that is symmetric and only serves to mix lineages among demes. Hence, the symmetric mixing fraction of relocation also equals the migration from sink to source.

In a structured scenario with a covariance of 'c' (see figure 3.1) the fraction of the parameter space between 'c' and the maximal covariance, relative to the fraction outside the domain, corresponds to the fraction of migration that is symmetric and mixing. Hence the larger the covariance the smaller the symmetric mixing fraction of migration.

Equations (3.47), (3.48) and (3.49) are plotted against $Cov(B, D)$ and b in figure 3.2 and against $Cov(B, D)$ and d in figure 3.3 The domain shown in figure 3.1

constitutes the base plane in the figures 3.2 and 3.3. The scenarios range from symmetry in net productivity to the case where the source has a net production 2.3 times that of the sink. This is well within the limits of biological realism. In each figure, the graphs for each mode of sampling are much alike. Only, there is a spike in the corner, when sampling evenly among demes. There should be a thin spike in the graphs for sink sampling as well, but the resolution of the graph fails to show it. In both figures the graphs fall off with larger covariance. Note that the surfaces spans only the domain given in figure 3.1, and not the entire square spanned by the two axises.

3.3.1 Effect of Relocation Waiting Time

This effect is described in section 2.2, and is responsible for the spike in the corner of the graphs for even sampling in the figures 3.2 and 3.3. The expected coalescence time may be greatly prolonged if the lineages are located in separate demes, and the relocation probabilities are low. In this case also cross coalescences will be rare. Hence, for very low relocation probabilities, lineages in separate demes will effectively be precluded from coalescing. The effect is obviously only in play for relocation probabilities so low, that it can not be assumed that the amount of time the lineages spend in the two demes is stationary distributed. This assumption is treated in section 3.4. Hence, the effect is only seen in the corner of the parameter space corresponding to minimal demographic difference between demes and maximal structure. That is, where relocation probabilities are very low. The effect is manifested in the results for $E(1,1)$ (even sampling) and $E(2,0)$ (sampling in the sink), by the increase in expected coalescence time.

The effect is most pronounced for even sampling, since, in any case, there will be at least some relocation waiting time. In the case of sink sampling the effect is less strong (there is actually a thin spike in the corner, although the figures fail to show it), since the possibility of a coalescence event does not rely entirely on a relocation event. The effect that is seen owes to cases where a lineage relocates before the two lineages coalesce, so that the coalescing of the two becomes dependent on another relocation event. This will of cause happen with larger probability the larger the relocation probability is. However, the resulting waiting time after the first relocation event decreases with larger relocation probability. Hence, as seen in the graph, we expect the mean of the coalescence time to grow as the relocation probability falls. For this mode of sampling, the expected coalescence time tends to infinity as the relocation probability tends to zero, but for a relocation probability equal to zero, the expected coalescence time is not infinite. Rather, for this value the expected coalescence time is equal to b, since in this situation all coalescences will happen in the sink without interference of migration.

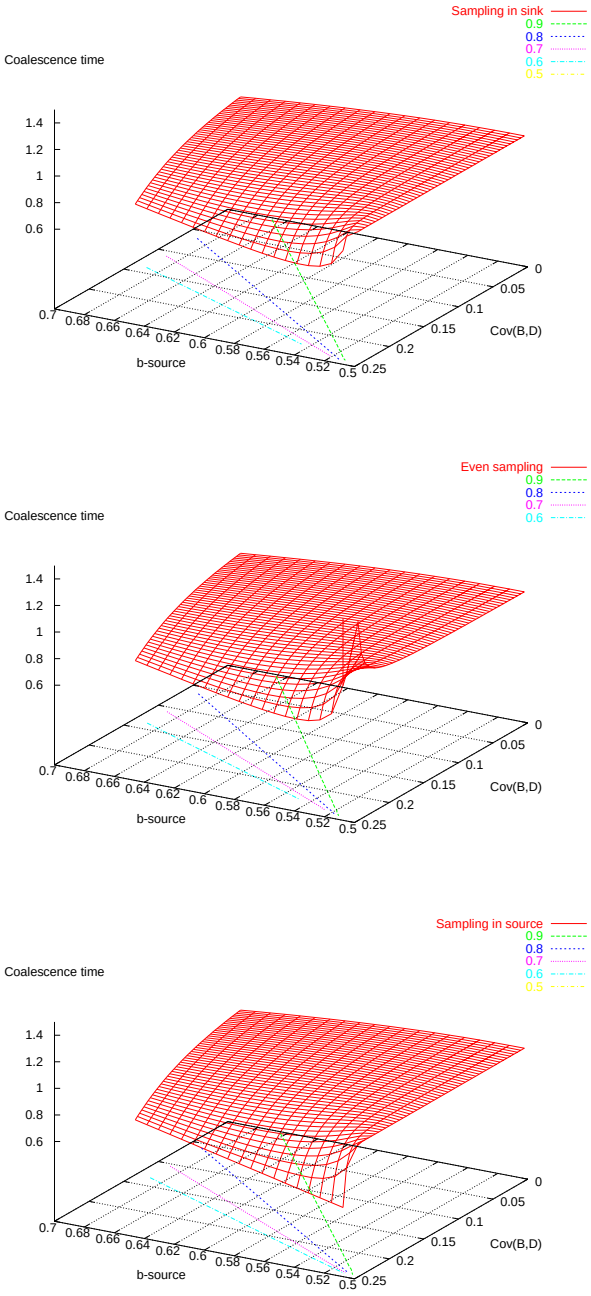


Figure 3.2: Expected coalescence time for two sequences in the death model. The two deme sizes are equal $N = 100$. The modes of sampling are from top to bottom: Two from the sink, one from each deme, two from the source.

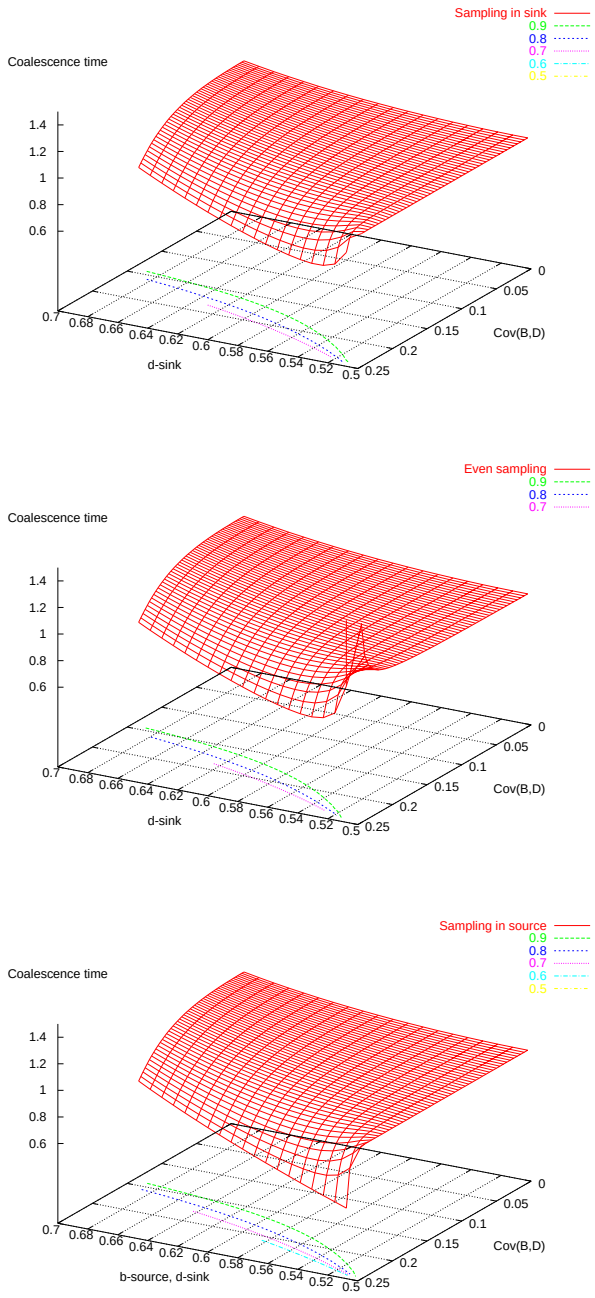


Figure 3.3: Expected coalescence time for two sequences in the birth model. The two deme sizes are equal $N = 100$. The modes of sampling are from top to bottom: Two from the sink, one from each deme, two from the source.

3.3.2 Effect of Asymmetric Relocation Probabilities

The asymmetry of the relocation probability results in an aggregation effect as described in section 2.2. The lineages will have a tendency to be located together in the source a larger fraction of the time, than they would with symmetric relocation probabilities. When there is an elevated probability of finding the lineages together in some smaller subsection of the population they obviously have a higher probability of finding a common ancestor each generation.

The covariance works to accentuate this effect. With a covariance equal to zero we have a unstructured scenario, where the probability of finding a lineage in the source is not much larger relative to finding it in the sink. With a maximal covariance, which means that only the relocation probability from sink to source is positive, the lineages will aggregate in the source. This is the why the expected coalescence time falls of towards maximal covariance. The magnitude of this effect is in part determined by the difference in the primary parameter values between the demes. The larger the difference, the more strongly asymmetric the relocation probabilities are, and the smaller the covariance has to be to even them out. In other words, the stronger the source-sink functionality, the more symmetric mixing migration is needed to counteract the effect of aggregation.

3.3.3 Effect of Local Demography Differences

This effect is the one responsible for the differences between the graphs in figure 3.2 and those in figure 3.3. A comparison of the graphs in figure 3.2 and 3.3 is shown in figure 3.4. The effect is a result of different in-deme coalescence rate between the two demes. With varying birth rates among demes, the in-deme coalescence rate in the source is larger than in the sink (since b is larger in the source), whereas with with varying death rates it is the other way around (since d is smaller in the source). The contour lines in figure 3.4 depict the differences in migration scenario for the same mean coalescence times in each model. The difference in the expected coalescence times for each parameter set is a consequence of different drift regimes through the history of the sample. In other words, the difference depends on how much of the time the lineages spend under which drift regimes.

For maximal structure (maximal covariance), the coalescence rate is effectively equal to the in-deme coalescence rate in the source if the relocation probability is just moderately strong. This is because only the relocation probability from sink to source is positive and at most two relocation events will occur. Recall that in the maximally structured situation the in-deme coalescence rate in the source is governed by d. The expected coalescence time for even sampling and sink sampling are affected by relocation waiting time for low relocation probabilities, and converges to $0.5^2/d$ (the inverse in-deme coalescence rate in the source), as the distribution of lineages among demes approaches stationarity. In the maximally structured situation stationarity means that the lineages spend effectively all their time in the source. For two sequences sampled in the source, the expected coales-

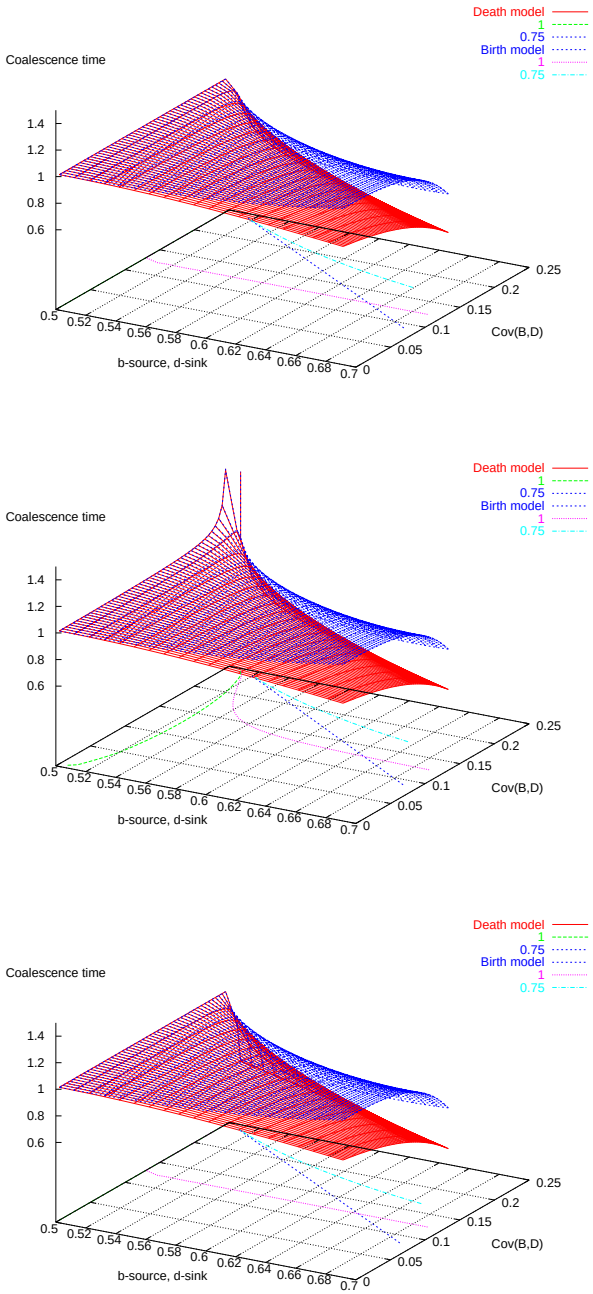


Figure 3.4: Expected coalescence times for two sequences in the death and the birth model. The modes of sampling are from top to bottom: Two from the sink, one from each deme, two from the source. The contour lines shows which migration scenarios that give the same mean coalescence time in the two models. $N = 100$.

cence time is of cause $0.5^2/d$ for the entire range of the primary parameter, since in this case there is obviously no relocation events. In conclusion, the expected coalescence time for the maximally structured situation of cause varies as a function of d in the birth model and is constant at 0,5 for the entire range in the death model.

For an unstructured setting (covariance zero), the in-deme coalescence rates for the two models become more different as the difference in the primary parameter increases. This is because the relocation probabilities become more asymmetric, so that the lineages are more often found in the source, and less often in the sink. Hence, the total coalescence rate is more affected by the in-deme coalescence rate in the source than that in the sink. In the death model, the in-deme coalescence rate is higher in the source and lower in the sink. This means that as the difference in the values of b among the demes gets larger, the expected coalescence time becomes smaller. In the birth model the in-deme coalescence rates are affected reciprocally. Hence, in this case expected coalescence time gets larger, the larger the differences in the values of d.

Figure 3.5 shows the difference in the expected coalescence time between the two models for the same migration scenarios. The difference for each mode of sampling is, not surprisingly, independent of relocation waiting time. For some covariance, the difference is a function of the difference in the value of the primary parameter between the two demes. The dependence on covariance for some set of birth and death rates, however, is not unambiguous. For small covariances the mixing migration is strong, resulting in a more uniform distribution of the lineages over the demes. Hence, the different drift regimes in the source and the sink, will chancel out to some extent. As the covariance grows the difference in figure 3.5 gets larger. This is due to the fact that as the population becomes more structured, the relocation probability corresponding to symmetric mixing migration becomes smaller. Hence, as the conserving one-way migration becomes more dominant, the lineages will spend more time in the source so that the drift regime in the source will dominate over the drift regime in the sink. In other words, the antagonistic effect of the drift regime in the sink will decrease. In conclusion, as the covariance gets larger, the expected coalescence time becomes more dependent on the drift regime in the source.

For large covariances the difference between the models falls of. This owes to the fact that, as the covariance approaches its maximal value, only the drift regime in the source *or* the sink is dependent on migration, depending on which parameter is the primary one (see (3.13)). This means that for maximal structure, the drift regime in the source is only affected by the primary parameter in the birth model, and the drift regime in the sink is only affected by the primary parameter in the death model. Hence, for some value of the primary parameter, the difference between the models, will decrease towards maximal covariance.

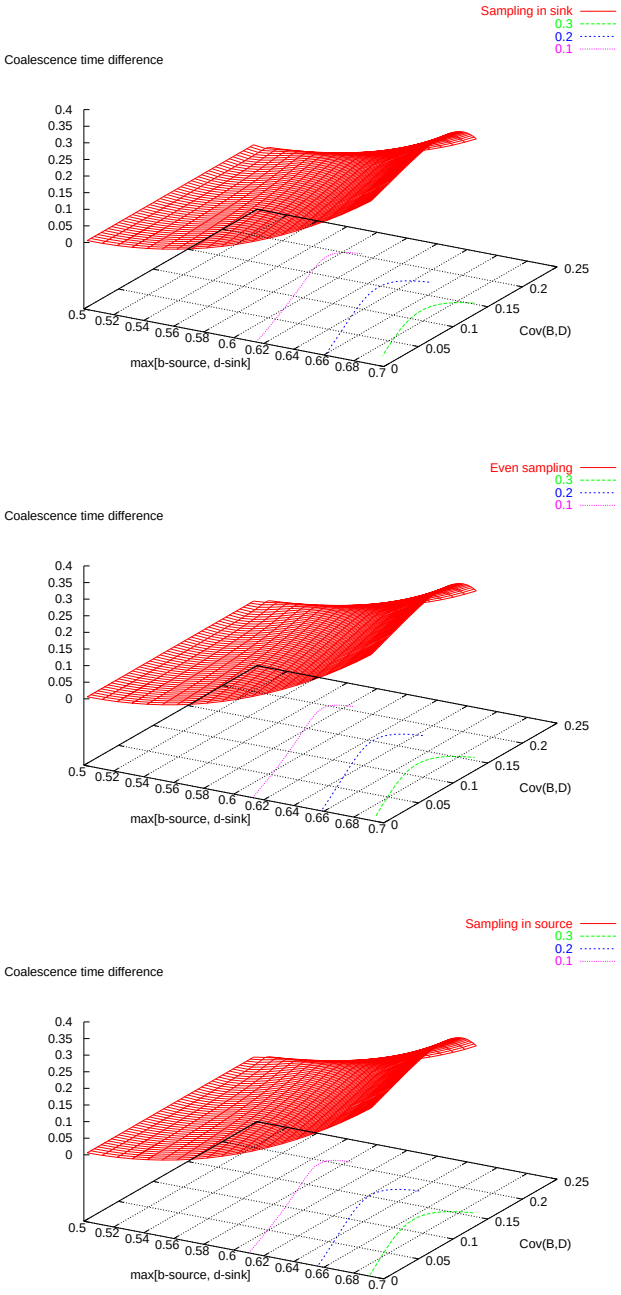


Figure 3.5: The deviation of the expected coalescence time for two sequences in the birth model from that in the death model. The modes of sampling are from top to bottom: Two from the sink, one from each deme, two from the source. $N = 100$

The Model-Specific Effects of Demography

The effect that local demography differences impose on expected coalescence time under the birth or a death model can be investigated by comparing the models to a reference with no demography differences between demes. Such a model is produced from the present model by setting *both* the b's and the d's in the in-deme coalescence rate terms to 0.5.

$$\frac{\binom{\alpha_i}{2}(0.5\ 0.5 + Cov(B, D))\varphi}{{c_i}^2} \quad i \in \{1, 2\}. \tag{3.50}$$

The emerging model can not meaningfully be explained from from the elements of the transition probabilities, and is only to be thought of as a reference model with a set of transition probabilities equal to what would be found in a setting with the same relocation probabilities but with no demography variation.

Figure 3.6 depicts the deviation in expected coalescence time in the death model, from the result obtained from the reference model neglecting the local drift differences, and Figure 3.7 shows this deviation for the birth model. The deviation due to demography differences are stronger in the birth model. However, compared to the reference model the deviations of the coalescence rates in the death and the birth model are equal. The difference between figure 3.6 and figure 3.7 owes to the fact that these plot the inverse coalescence rates, i.e. the expected coalescence times.

The drift differences in each will together with the deme sizes, comprise the effective deme sizes that may be inferred from a data set. The deme sizes that would be obtained from a population described by the model presented here, would thus be the deme sizes resulting in the same strength of drift in each deme, as described above.

Effective Population Size Description

For deme sizes of 100, as shown in figure 3.2 and 3.3, the expected coalescence time shows only a slight dependence on the mode of sampling in most of the parameter space. In the corner of the parameter space where the relocation rates are so low that initial sampling may potentially play a role, the difference in in-deme coalescence rate is very small, as seen in figure 3.5. This implies that for combinations of parameter values that produce a local drift effect due to varying demography the process can be approximated by the strong migration limit (see section 2.4.4) as long as the deme sizes are just moderately large ($N > 100$).

As will be considered in section 3.4.2, stationarity of the distribution of lineages among demes, implies that the Coalescent of the sample is a standard Kingman one, which again implies that the effect of a source-sink functionality on the expected coalescence time is an effect on effective population size only. Hence, the population behaves like an unstructured population with different effective total

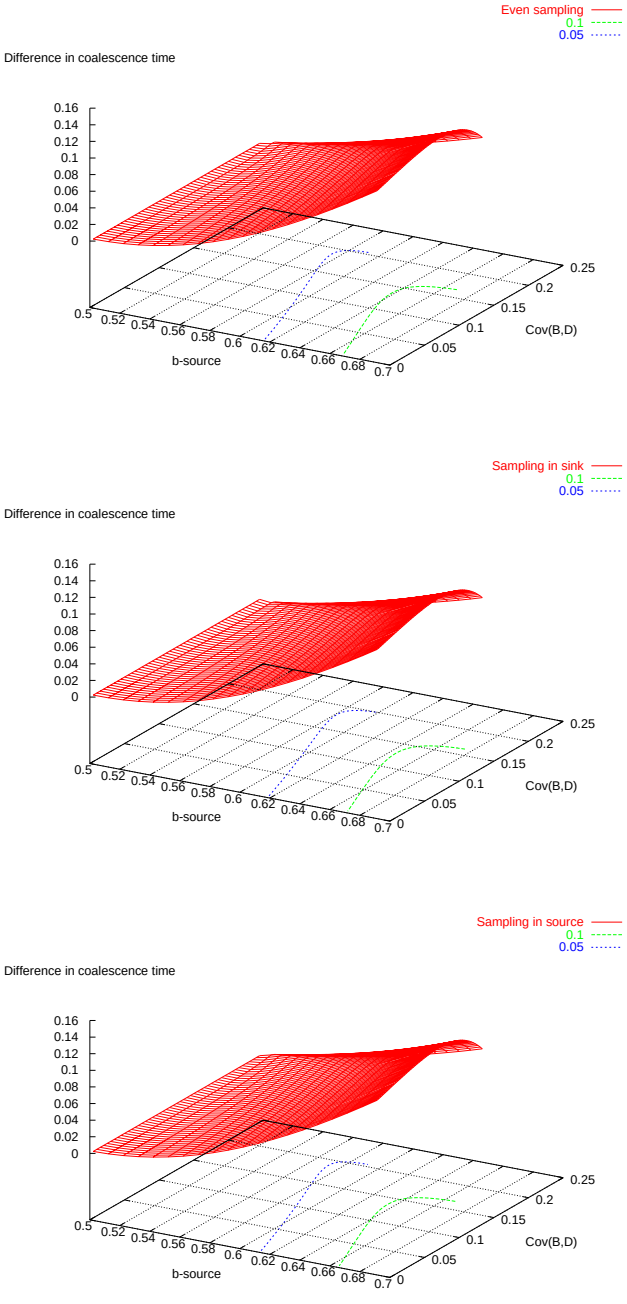


Figure 3.6: The deviation of the expected coalescence time of two sequences in a model with no local demography differences, from that in the death model. The modes of sampling are from top to bottom: One from each deme, two from the sink, two from the source. $N = 100$. The expected coalescence time for no demography differences were obtained from equation (3.47), (3.48) and (3.49), setting the in-deme coalescence rates to (3.50).

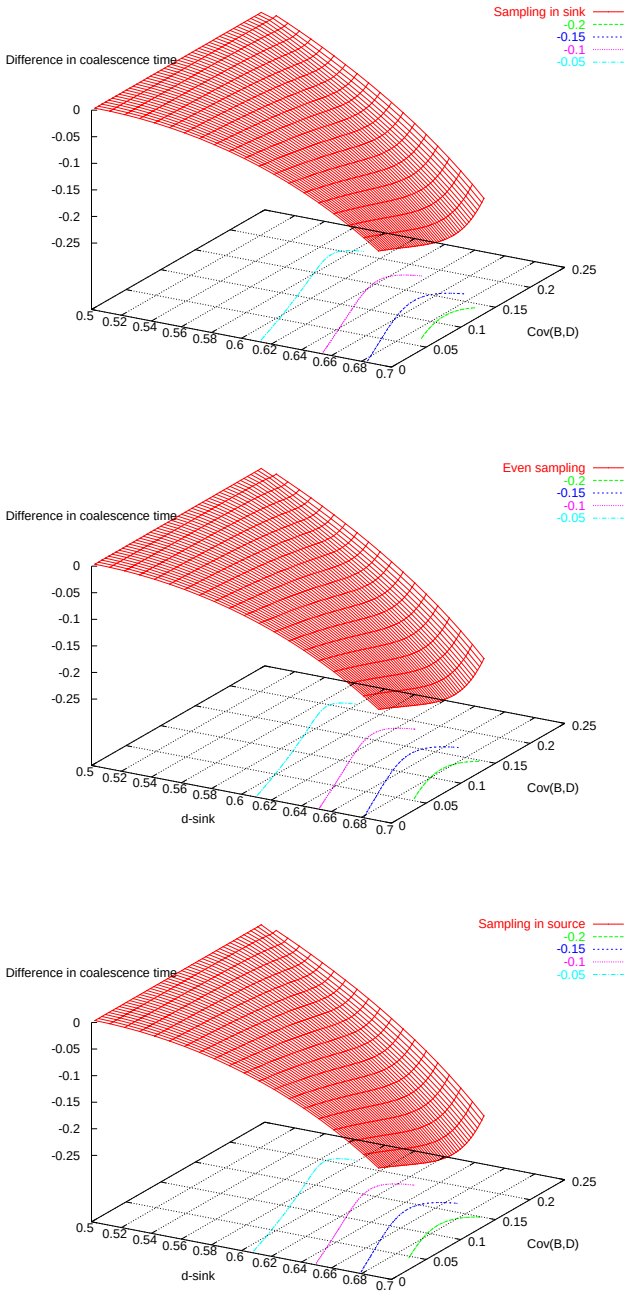


Figure 3.7: The deviation of the expected coalescence time of two sequences in a model with no local demography differences, from that in the birth model. The modes of sampling are from top to bottom: One from each deme, two from the sink, two from the source. $N = 100$. The expected coalescence time for no demography differences were obtained from equation (3.47), (3.48) and (3.49), setting the in-deme coalescence rates to (3.50).

population size for parameter values that give an effect of a source-sink dynamic, as long as the population size is just moderately large.

The source-sink effective population is described in detail in section 3.4. For deme sizes so small that the process can not be approximated by the strong migration limit, the demographic effects may have an effect on tree structure. This is considered in the next section.

3.3.4 Effects on Tree Structure

Before we go on to describe the source-sink effective population size that was found to be a sufficient description for just moderately large deme sizes, the effects on tree structure for small demes is investigated.

For a sample of more than two lineages, the topology and relative branch lengths of the resulting trees becomes interesting if these deviate from what would be expected for asymmetric relocation probabilities but no difference in demography among demes. This would be the case if demography changes the effective sizes of demes. Hence, the effect of local demography differences would not only be an effect on total effective population size, but also an effect on the effective deme sizes. That is, not only an effect on N_e but also an effect on $c = \{c_1 \dots c_D\}$.

Deviations from a situation without drift difference will be manifested in the tree structure if the lineages are subject to different strengths of drift at different periods of time. This would be the case if relocation is unidirectional or highly asymmetric, and the relocation probabilities are of the same magnitude of the coalescence rates. In this case, sampling from the sink will produce a tree where the drift regime regime near the present differs from that near the root of the tree. If the relocation probability from sink to source is much larger than the coalescence rate in the sink, the sampled lineages will spend all their time in the source. Hence, the drift regime will not differ over the tree.

If the strength of drift near present time, where the lineages are resident in the sink, is different from the strength of drift further back in time where the lineages reside in the source, due to the difference in in-deme coalescence rate in the demes, the tree structures are expected to differ between the death and the birth model. In the death model coalescence rate is expected to be slower near present time, and faster near the common ancestor. In the birth model it should be the other way around. In the maximally structured situation, the in-deme coalescence rate and the relocation probability in the sink are equal for k lineages if sink deme size is given by

$$N_1 = \frac{(k-1)b_1}{b_2 d_1 - b_1 d_2}, \tag{3.51}$$

where 1 is the sink and 2 is the source. The larger the difference in the the value of the primary variable, the smaller the deme size have to be. Assuming equal deme sizes, these examples of parameter sets make the rates in (**??**) equal when the number of lineages in the sink is seven.

Primary parameter		Deme size in death Model	Deme size in birth Model
0.7	0.3	9	15
0.65	0.35	14	20
0.6	0.4	24	30
0.55	0.45	54	60

This implies that deviations in tree structure due to demography are to be looked for in this neighbourhood of parameter values.

The effect of source-sink functionalities on the ancestral relationship of a sample is particularly interesting for small populations, since it is often in the cases where population sizes are small, that the questions concerning the independent survival of a demes, are asked.

Simulation

I have written a simulation programme in C that generates coalescence times and tree statistics under a Moran model with specified parameters. It simulates a wide variety of scenarios. Adjustable parameters are per capita birth and death parameters of each deme, covariance of birth an death events, relative sizes of demes, total population size and mode of sampling. In outline the simulation algorithm is:

1. The time to the next event is sampled from an exponential distribution, with parameter equal to the sum of all relocation and coalescence probabilities.

2. It is determined whether the next event it is a coalescence event or a relocation event by a simple weighting of exponential intensities.

3. It is determined which deme/demes are affected.

4. In the case of a relocation event, it is determined whether the relocation is actually a cross coalescence event, and if it is, which two lineages that coalesce.

5. In the case of a in-deme coalescence event, is is determined which two lineages that coalesce.

6. 1 through 5 is repeated until only one lineage remains.

7. Branch lengths and tree statistics are calculated.

8. 1 through 7 are repeated 1000000 times.

9. Means and variances of of branch lengths and tree statistics are returned to output.

The means are obtained as the arithmetic mean of the results of each of the 1000000 runs, since each result occurs according to the density under the simulated model.

The simulations were done to investigate the effects of a source-sink functionality on tree structure. The maximally structured death and birth model are studied in the cases where they produce the same source-sink functionality. That is, when the values of b_1 and b_2 in the death model equals the values of d_2 and d_1 in the birth model. Hence, the relocation regimes are the same, whereas the drift regimes differ between the models. For obvious reasons, it is not possible to compare the models in a situation where the drift regimes are the same for both models. This would result in symmetric relocation probabilities, and thus obliterate the source-sink relation under study.

In addition, results are simulated for the same unidirectional migration regime, but artificially neglecting the demography specific differences in in-deme coalescence rate (see section 3.50). This is intended as a reference to separate the effects of pure asymmetric migration from the effects of demography specific drift differences. To get the clearest picture possible, deme sizes are set to be even.

The sample size is eight. For each set of values of the primary parameters, a deme size is chosen, so that the coalescence rate approximately equals the relocation probability in the sink, when seven lineages remain in the sink. Each of the considered cases simulated for three modes of sampling: Sampling entirely from the sink, from the source, or evenly from both sink and source. Apart from the expected coalescence times, the following tree statistics are obtained:

T: Tree Depth, the time to the most recent common ancestor.
A: Total branch length, the sum of the length of all the branches.
E: External branch length, the sum of the terminal branches.
I: Total internal branch length, the sum of the non-terminal branches.
L: Last branch level, the coalescence time of the two last lineages.

As described in the previous chapter the values of these statistics under the standard Kingman Coalescent, together with the values for $n = 8$ are:

E(T) = $2(1 - 1/n)$ =1.75
E(A) = $2\sum_{i=1}^{n-1} 1/i$ =5.186
E(E) = 2 =2
E(I) = $E(A - E)$ =3.186
E(L) = 1 =1
E(E)/E(A) = 0.387
E(E)/E(T) = 1.143,

(Kingman 1982*a*),

Simulation Results

The results are summarised in table 3.1 and 3.2. The results in table 3.1 are for primary parameters 0.7 and 0.3, and those in table 3.2 are for 0.6 and 0.4.

Recall that in the maximally structured setting, the in-deme coalescence rate in the sink is dependent on the birth rate only, and that the coalescence rate in the source is dependent on the death rate only, (see (3.13) for reference). Hence, in the death model, the drift regime will be weaker in the sink corresponding to a larger deme size, whereas the drift regime in the source will be unaffected. In the birth model the drift regime will be weaker in the source and unaffected in the sink.

The deviations from the reference case for per capita parameters 1.2 and 0.8 are rather small. Hence, the values of the per capita parameters must differ at least as much as 1.4 and 0.6 if demographic differences between the demes is to have a considerable influence on tree structure.

Sampling from the source gives trivial results since such a sample will be unaffected by migration. Hence a standard Kingman Coalescent results. Interest focuses on the sink sample that is subject to different drift regimes back through time. The expected coalescence times in the death and the birth model reflects the history of changing drift regimes. In the death model, the first branch levels are longer and the last are shorter, as expected. The opposite effect is seen for the birth model. The relation between initial coalescence rate and relocation probability are not the same for the death and the birth model. Recall that the deme sizes that makes the coalescence rate and the relocation probability equal are not the same for both models (see section 3.3.4). For the same relocation probabilities, the initial coalescence rate in the sink is higher in the birth model than in the death model. In the reference case the relation between the rates equals that in the birth model.

Comparing the tree statistics from the death and the birth model present the following picture: The total tree depth is much larger in the birth model. This is because the last coalescences with the longest waiting times, take place in the source, where the drift regime is weakest in the birth model. The total branch length is only slightly larger in the birth model. This implies that the trees in the death model must have longer first branch levels to compensate for the longer last branch levels in the death model.

It might be expected that longer first branch levels would result in longer external branch length. However, both the external branch lengths and the ratios external/total branch length, show only a slight difference between the models. The ratio external branch length/total tree depth, however, is much larger in the death model. This may result from different topologies occurring with different probabilities in the two models:

In the birth model the coalescence rate in the sink relative to the relocation probability is higher than in the death model (see the table in section 3.3.4). Hence, it is more probable that lineages will coalesce before relocation than in the death model. If a lineage do relocate before it is involved in a coalescence, it is more probable in the birth model that only a few lineages will join it later, since most of

the lineages in the sample will have coalesced in the sink. Hence, such a lineage will have a larger probability of coalescing as one of the last in the sample, that is, representing an external branch in the entire length of the tree.

In conclusion, the birth model may have the same relative fraction of external branch length as the death model, but the external branches are distributed differently in the trees. In the death model, it is more probable that the first branch level are long but less probable that the external branches extend far back in the tree. In the birth model, it is less probable that the first branch level are long but more probable that the external branches extend far back in the tree.

The results for sink sampling and even sampling are much alike differing only in the first branch levels. Here the expected coalescence time is slightly longer for even sampling due to an effect of coalescence preclusion.

3.4 Strong Migration Approximation

As shown in section 3.3.3, the effects of a source-sink functionality between demes, can be described as an effect on effective population size only in most of the parameter space, as long as the deme sizes are just moderately large. This is because the relocation probabilities are so large relative to the total coalescence rate, that the distribution of lineages among the demes is effectively stationary. Refer to section 2.4.4 for a description of the strong migration limit.

For the approximation to be justified relocation probabilities must be so much larger than the total coalescence rate, that a stationary distribution of lineages among demes between coalescence events can be assumed. Hence, the crucial relation is R_{i*}/λ_k, which must be large for all i. R_{i*} is the scaled probability of relocating from deme i, and λ_k is the scaled coalescence rate for k remaining lineages, calculated under the assumption of strong migration. The reliability of the assumption will be considered separately.

Recall that the ancestry in populations in the strong migration limit is described by a standard Kingman Coalescent with a migration effective population size. This implies that the ancestral relationship of sequences sampled from a large source-sink population is described simply by a scaling of the standard Kingman Coalescent. In this section, results for this new source-sink effective population size, N_e, will be presented.

3.4.1 The Source-sink Effective Population Size

Since a standard Kingman Coalescent is assumed, so that the relative lengths of branch levels are known, we only need to consider the results for two sequences. The stationary distribution is binomial for two sequences and gives the probability that l of the two lineages are located in deme one

$$p(l) = \binom{2}{l} \pi_1^l (1-\pi_1)^{2-l}, \tag{3.52}$$

Reference Case:

Sampling	Sink	Source	Even
E(8)	0.0286 ± 0.0333	0.0247 ± 0.0278	0.0414 ± 0.0441
E(7)	0.0441 ± 0.0441	0.0330 ± 0.0330	0.0534 ± 0.0534
E(6)	0.0686 ± 0.0704	0.0461 ± 0.0479	0.0716 ± 0.0729
E(5)	0.1065 ± 0.1065	0.0694 ± 0.0694	0.1005 ± 0.1005
E(4)	0.1673 ± 0.1681	0.1156 ± 0.1161	0.1529 ± 0.1540
E(3)	0.2883 ± 0.2883	0.2313 ± 0.2313	0.2698 ± 0.2698
E(2)	0.7276 ± 0.7277	0.6948 ± 0.6950	0.7149 ± 0.7150
E(T)	1.4311	1.2153	1.4048
E(A)	4.4718	3.6001	4.4897
E(E)	1.8525	1.3879	2.0018
E(I)	2.6192	2.2121	2.4878
E(L)	0.7276	0.6948	0.7149
E(E)/E(A)	0.4142	0.3855	0.4458
E(E)/E(T)	1.2944	1.1420	1.4249

Death Model:

Sampling	Sink	Source	Even
E(8)	0.0373 ± 0.0394	0.0198 ± 0.0262	0.0402 ± 0.0437
E(7)	0.0535 ± 0.0535	0.0264 ± 0.0264	0.0507 ± 0.0507
E(6)	0.0754 ± 0.0770	0.0370 ± 0.0404	0.0663 ± 0.0687
E(5)	0.1051 ± 0.1051	0.0555 ± 0.0555	0.0912 ± 0.0912
E(4)	0.1516 ± 0.1522	0.0926 ± 0.0936	0.1355 ± 0.1363
E(3)	0.2479 ± 0.2479	0.1848 ± 0.1848	0.2332 ± 0.2332
E(2)	0.5972 ± 0.5972	0.5551 ± 0.5553	0.5873 ± 0.5874
E(T)	1.2683	0.9716	1.2048
E(A)	4.1968	2.8798	3.9486
E(E)	1.9271	1.1110	1.8829
E(I)	2.2697	1.7688	2.0657
E(L)	0.5972	0.5551	0.5873
E(E)/E(A)	0.4591	0.3857	0.4768
E(E)/E(T)	1.5194	1.1434	1.5628

Birth Model

Sampling	Sink	Source	Even
E(8)	0.0230 ± 0.0292	0.0330 ± 0.0363	0.0429 ± 0.0465
E(7)	0.0369 ± 0.0369	0.0441 ± 0.0441	0.0584 ± 0.0584
E(6)	0.0624 ± 0.0647	0.0616 ± 0.0634	0.0813 ± 0.0844
E(5)	0.1081 ± 0.1081	0.0926 ± 0.0926	0.1178 ± 0.1178
E(4)	0.1889 ± 0.1895	0.1543 ± 0.1547	0.1835 ± 0.1842
E(3)	0.3525 ± 0.3525	0.3082 ± 0.3082	0.3343 ± 0.3343
E(2)	0.9500 ± 0.9501	0.9257 ± 0.9258	0.9370 ± 0.9371
E(T)	1.7220	1.6197	1.7555
E(A)	5.0714	4.8000	5.4414
E(E)	1.8935	1.8521	2.2640
E(I)	3.1779	2.9478	3.1774
E(L)	0.9500	0.9257	0.9370

Reference Case:

Sampling	Sink	Source	Even
E(8)	0.0254 ± 0.0316	0.0206 ± 0.0242	0.0397 ± 0.0429
E(7)	0.0408 ± 0.0408	0.0275 ± 0.0275	0.0496 ± 0.0496
E(6)	0.0640 ± 0.0659	0.0385 ± 0.0402	0.0646 ± 0.0667
E(5)	0.0965 ± 0.0965	0.0579 ± 0.0579	0.0879 ± 0.0879
E(4)	0.1457 ± 0.1467	0.0965 ± 0.0970	0.1304 ± 0.1315
E(3)	0.2416 ± 0.2416	0.1929 ± 0.1929	0.2258 ± 0.2258
E(2)	0.6054 ± 0.6056	0.5792 ± 0.5792	0.5957 ± 0.5958
E(T)	1.2197	1.0134	1.1939
E(A)	3.8754	3.0029	3.8836
E(E)	1.6471	1.1571	1.7824
E(I)	2.2283	1.8457	2.1012
E(L)	0.6054	0.5792	0.5957
E(E)/E(A)	0.4250	0.3853	0.4589
E(E)/E(T)	1.3503	1.1417	1.4928

Death Model

Sampling	Sink	Source	Even
E(8)	0.0284 ± 0.0312	0.0185 ± 0.0225	0.0388 ± 0.0424
E(7)	0.0442 ± 0.0442	0.0247 ± 0.0247	0.0480 ± 0.0480
E(6)	0.0659 ± 0.0671	0.0347 ± 0.0374	0.0616 ± 0.0634
E(5)	0.0948 ± 0.0948	0.0520 ± 0.0520	0.0835 ± 0.0835
E(4)	0.1376 ± 0.1384	0.0868 ± 0.0875	0.1226 ± 0.1236
E(3)	0.2240 ± 0.2240	0.1737 ± 0.1737	0.2097 ± 0.2097
E(2)	0.5500 ± 0.5501	0.5209 ± 0.5211	0.5414 ± 0.5415
E(T)	1.1452	0.9116	1.1058
E(A)	3.7302	2.7012	3.6372
E(E)	1.6553	1.0418	1.7168
E(I)	2.0748	1.6593	1.9203
E(L)	0.5500	0.5209	0.5414
E(E)/E(A)	0.4437	0.3857	0.4720
E(E)/E(T)	1.4454	1.1428	1.5524

Birth Model

Sampling	Sink	Source	Even
E(8)	0.0228 ± 0.0261	0.0232 ± 0.0259	0.0406 ± 0.0430
E(7)	0.0377 ± 0.0377	0.0309 ± 0.0309	0.0519 ± 0.0519
E(6)	0.0622 ± 0.0640	0.0434 ± 0.0448	0.0682 ± 0.0698
E(5)	0.0987 ± 0.0987	0.0650 ± 0.0650	0.0939 ± 0.0939
E(4)	0.1546 ± 0.1554	0.1086 ± 0.1093	0.1400 ± 0.1408
E(3)	0.2636 ± 0.2636	0.2166 ± 0.2166	0.2462 ± 0.2462
E(2)	0.6755 ± 0.6756	0.6498 ± 0.6499	0.6651 ± 0.6653
E(T)	1.3153	1.1378	1.3063
E(A)	4.0746	3.3727	4.1977
E(E)	1.6565	1.3012	1.8729
E(I)	2.4181	2.0715	2.3247
E(L)	0.6755	0.6498	0.6651

where π_i is the probability of finding one lineage in deme i, or the fraction of the time, that one lineage spends in deme i. The vector π can be can be obtained as the left eigen vector with corresponding eigen value one, of the backwards migration matrix $\mathcal{R}$.

$$\mathcal{R} = \begin{bmatrix} 1 - r_{12} & r_{12} \\ r_{21} & 1 - r_{21} \end{bmatrix}, \tag{3.53}$$

and is given by

$$\pi_j = \frac{r_{ij}}{r_{ij} + r_{ji}} \quad j \neq i \;\; i, j \in \{1, 2\}, \tag{3.54}$$

where

$$r_{ij} = \frac{(d_i b_j - Cov(B, D))(c_j N_T - 1)}{2 c_i c_j}. \tag{3.55}$$

$p_i(l)$ can be interpreted both as the probability of finding l lineages in deme i, and the expected fraction of the time that l lineages reside in deme i. This means that the rate of coalescence for the two sequences, λ_2, can then be calculated as:

$$\begin{aligned} \lambda_2 \;\; = \;\; & p(2) \times \frac{b_2 d_2 + Cov(B, D)}{c_2^2} + p(0) \times \frac{b_1 d_1 + Cov(B, D)}{c_1^2} \\ & + p(1) \times \left(\frac{b_1 d_2 - Cov(B, D)}{2 c_1 c_2} + \frac{b_2 d_1 - Cov(B, D)}{2 c_1 c_2} \right). \end{aligned} \tag{3.56}$$

The source-sink effective population size is thus given by:

$$N_e = \frac{N_T}{\sigma^2 \lambda_2}. \tag{3.57}$$

In terms of the Wright-Fisher model for which effective population size is defined $\sigma^2 = 1$ and in the Moran model $\sigma^2 = 2/N_T$. This straightforward conversion is possible because the two models are exchangeable in the sense described in section 1.2. For reference, recall that $N_e = N_T/\sigma^2$ in a panmictic population with isotropic strong migration and uniform drift regimes in all demes.

In figure 3.8 N_e, in units of $N^2/2$, is plotted as a function of $Cov(B, D)$ and either b or d as primary parameter. Note that it is effectively identical to the graphs in figure 3.4 except for small difference in the values of the primary parameter and large covariance. Figure 3.9 shows the difference in N_e, in units of $N^2/2$, of both the death and the birth model to a reference model with no local differences in demography. Note that the effect of asymmetric relocation probabilities that tend to aggregate lineages in one deme thus lowering the expected coalescence time is included in the reference model. Hence, figure 3.9 shows only the impact, of local demographic differences in a source-sink population, on the effective population size. As explained in section 3.3.3, the deviation due to demography differences

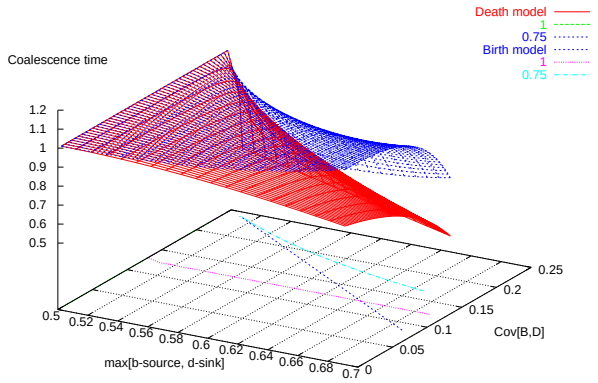


Figure 3.8: The expected coalescence time, N_e, for two sequences is plotted as a function of covariance and the largest value of the primary parameter, for both the death and the birth model.

are stronger in the birth model. However, the deviations in coalescence rates in the death and the birth model are equal. The difference between figure 3.6 and figure 3.7 owes to the fact that these plot the inverse coalescence rates, i.e. the expected coalescence times.

Since we have a standard Kingman Coalescent in the strong migration limit, the relative lengths of expected coalescence times are given by $1/\binom{k}{2}$, Hence, knowing the coalescence rate for two sequences, λ_2, the expected coalescence time of k is simply obtained as

$$E[T_{k\to k-1}] = 1 \Big/ \lambda_2 \binom{k}{2}. \tag{3.58}$$

3.4.2 Robustness of the Strong Migration Approximation

For the strong migration approximation to hold the relation R_{i*}/λ_k must be large for all i. Otherwise lineages may be "trapped" in some demes, thus preventing the very large number of relocation events needed, if the assumption of stationarity of the distribution of lineages among demes, is to be valid. There is a special case, however, where the spatial distribution is stationary because the lineages spend all their time in *one* deme, as in the case of nearly maximal structuring. In this case the approximation may be valid even though the probability of relocation from the source to the sink is nearly zero, because such an event would immediately result in a relocation back into the source. Hence, the lineages are not "trapped" in the sink.

In figure 3.10 and 3.11 the effect of assuming a stationary distribution of lineages is shown. The results generated under the stationary distribution under the

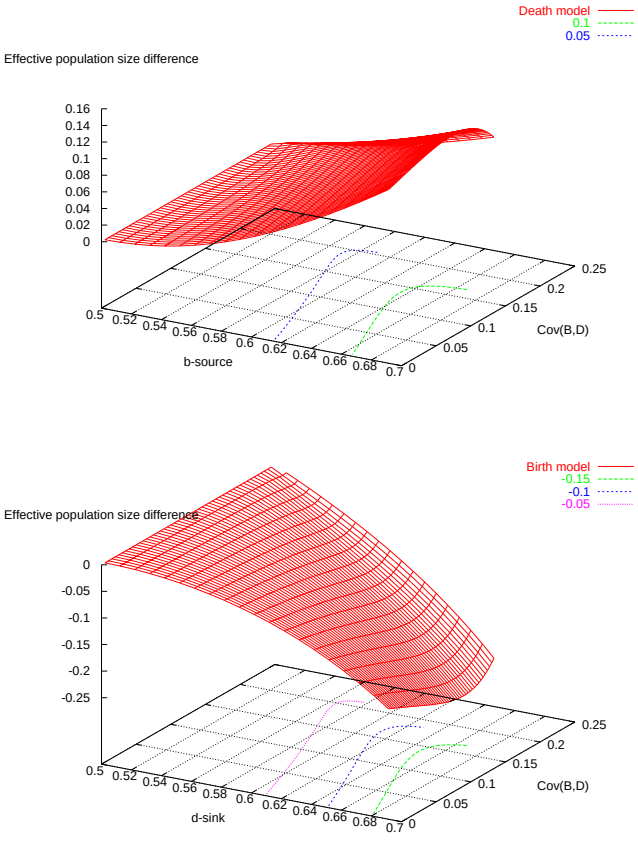


Figure 3.9: The difference in N_e induced by differences in demography among demes. The graphs show the deviation of N_e in a model with no local demography differences, from that in the death model (top) and the birth model (bottom). $N = 100$

strong migration approximation are compared to the exact results obtained from (3.47), (3.48) and (3.49). This gives an impression of the parameter space covered by the approximation. The results shown is for two sequences. The peaks in the graphs for even and sink sampling are due to the fact that the exact results cover relocation waiting time, whereas the approximation does not. The holes in the graphs for source sampling, are due to the fact that the approximation does not take into account the effect of early coalescences that follows from a low relocation probability out of the source. The edges of the graph for source sampling are zero (the approximation is exact). This is because both a maximal covariance and no difference in primary parameter values among demes correspond to no relocations out of the source. Hence, both cases corresponds to a stationary distribution of lineages where the lineages are only located in the source.

The figure 3.10 and 3.11 only shows the reliability of the approximation in the case of two sequences. As the number of sequences sampled from the same deme grows, initial coalescence rate will grow quadraticly, where as the relocation probability will only grow linearly. The accuracy of the approximation relies on the relation R_{i*}/λ_k. Hence, if the sample size is ten, and sampling is from one deme, the coalescence rate will initially be $(10-1)/4 = 45$ times that for two sequences, whereas the relocation probability will only the be five times larger than the case for two sequences. This implies that if the approximation is to be as good as in figure 3.10 and 3.11 the deme sizes would have to be $45/5 = 9$ times as large, that is $N = 900$.

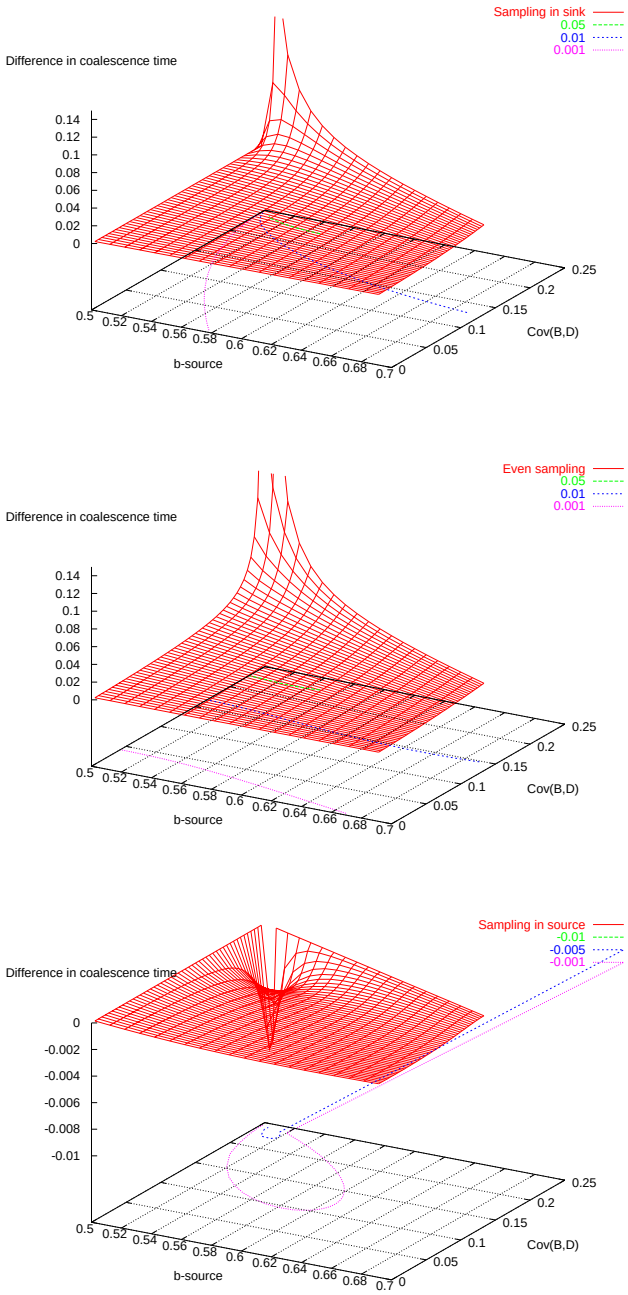


Figure 3.10: For the death model, the graphs show the deviation of the exact results obtained from (3.47), (3.48) and (3.49), from the results generated by the strong migration approximation. The modes of sampling are from top to bottom: Two from the sink, one from each deme, two from the source. $N = 200$.

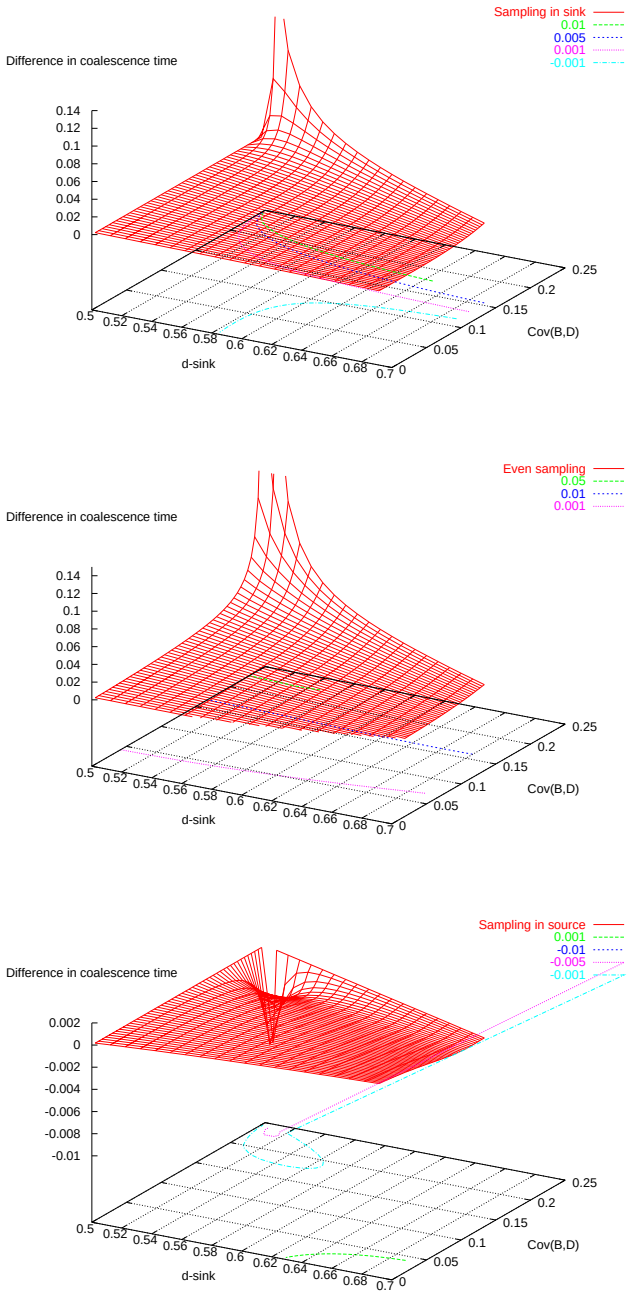


Figure 3.11: For the birth model, the graphs show the deviation of the exact results obtained from (3.47), (3.48) and (3.49), from the results generated by the strong migration approximation. The modes of sampling are from top to bottom: Two from the sink, one from each deme, two from the source. $N = 200$.

Chapter 4

Discussion

This chapter is divided into three parts. The first part is a discussion of the model presented in chapter 3. The next part is discussion of the information in the backwards migration matrix, and the last part is a discussion of some general problems in retrospective population genetic analysis.

4.1 Structured Moran Model

In retrospective genetic analysis it is crucial to distinguish between the parameters that can at most be obtained from a data set, and what can only be speculation as to how these parameters are produced. The only information that evolution leaves behind in the sequences, that may be sampled at present time, is the drift regime in each deme and the backward migration matrix. It is not possible to distinguish actual deme sizes from other factors influencing genetic drift, such as demography. Hence, the population size obtained from data is an effective population size. This implies that the maximal resolution of information is given by the vector $c = \{c_1 \dots c_D\}$ of scaled deme sizes, and the scaled relocation probabilities, $N_e r_{ij}$, composing the backwards migration matrix. Each deme size can be estimated as $c_i N_e$ but may hide all sorts of effects producing local drift differences. E.g. it is not possible to distinguish a source-sink functionality among even-sized demes, from a situation of plain asymmetric migration and different deme sizes. This composite nature of the effective deme sizes and effective population size, leaves a lot of space for interpretation. Hence, an understanding of the extent to which different effects may influence the deme sizes obtained from data, is of great value.

The Wright-Fisher model describes the backwards migration process and deme sizes down to these composite parameters, and is thus in line with what can maximally be obtained from data. Hence, if the relationships between effects, that may together produce effective deme sizes, is to be investigated, a more detailed description must be used.

4.1.1 Model

The Moran model formulated in chapter 3 describes the relationship between demographic effects, deme sizes and relocation probabilities in a population composed of a source and a sink that through migration upholds an equilibrium of deme sizes. This is done by expressing the coalescence rates and relocation probabilities as functions of birth and death rate in each deme, the covariance of birth and death events, the fractional sizes of demes, and the total population size.

The model assumes that the asymmetry of migration is a direct consequence of the demographic differences. This implies that the results derived in chapter 3 only applies to populations where a perfect distribution of surplus individuals among sinks is the case. The model may in principle describe gamete migration and haploid individual migration equally well, if it is assumed that both gametes and individuals disperse/migrate so that all demes may be reached, and migration on average will compensate for the difference in productivity. An example is gametes that move around among demes until a free space is found to settle in. This free space is found with a higher probability in a sink. However, passive dispersal of gametes will rarely conform to these assumptions. It is much more likely to be the case for individual migration, where an evaluation of habitat quality may be possible.

In the structured Moran model each relocation probability is totally resolved into the parameters b, d, c, $Cov(B, D)$ and N_T. This means that there is no degrees of freedom left in the relocation probabilities that can be used to make the backwards migration matrix reflect features of non-abstract geographical structure (that relocations between demes farther apart are less probable). This is possible in the W-F model since the relocation rates here are not completely explained.

4.1.2 Results

In the model presented here, the actual deme size, and the additional demographic effects due to local demography differences are separable in the sense that their individual effects may be investigated. However, the elements described in the model can not be separated through data analysis. The model only serve to add to an understanding of the effects of demography in source-sink model on effective deme sizes.

By an analytic approach it was found that the situations where an effect of demography may be seen, can be described by the strong migration approximation as long as deme sizes are moderately large. In section 3.4 the deviation of the effective population size arising from local drift differences due to demography was assessed. The cases investigated assume equal deme sizes, since interest focuses on the effects of demography differences and not those of different deme sizes. Hence, the death model implies even per capita death parameters, and the birth model implies even per capita birth parameters. The deviations in effective population size between the death and the birth model is almost 30%, in the case where the

per capita over/underproduction is 0.4. In other words, where the source deme produces 40% of the lineages in the sink deme. This is well within the limits of biological realism, and more extreme source-sink relationships are not improbable. The deviation, from the reference model with no drift difference in the demes, is in this case almost 20% for the birth model, almost 10% for the death model. Hence, the effective population size may be greatly influenced by demography. For larger differences in the values of per capita parameters, the difference in effect on N_e of the death and the birth model, will be even more pronounced. Note that the situation where a small part of the population may be responsible for the survival of the entire population does not necessarily result in a smaller effective population size, as one may think at first.

Through simulations it was concluded that local differences in demography may influence not only total effective population size, but also the effective deme sizes. That is, not only N_e but also $c = \{c_1 \dots c_D\}$. However, this will only be the case for very small deme sizes ($N < 20$), and for a pronounced source-sink relationship between demes (per capita parameters > 1.4 and < 0.6).

4.2 Inference from The Backward Migration Matrix

The backward migration matrix is a description of the genetic effect of migration. Each entry is given by

$$r_{ij} = \frac{c_j m_{ji}}{\sum_{k \in S} c_k m_{ki}}, \tag{4.1}$$

where r_{ij} is the relocation probability from i to j and m_{ji} is the forward migration probability from j to i. c_i refers to the fraction of N_T, that is, to the actual number of individuals in deme i. The denominator in (4.1) equals the fraction size of deme i, *after* the migration event, and *before* a possible regulation of deme size. This after migration fraction size is denoted c_i'.

By the simple rescaling applied section 2.1.3 to obtain the relocation rates from the migration rates it was assumed that the denominator in (4.2), c_i', equals c_i. This corresponds to assuming that, each migration event does not change the deme sizes. This, however, can not generally be assumed. In a Wright-Fisher model of finite size, migration will change the deme sizes at least to some extent. In the structured Moran model presented here, however, migration events does not change deme sizes.

To obtain the forward migration matrix from the backwards one we would have to solve a system of linear equations for each entry in the forward migration matrix:

$$m_{ji} = \frac{c_i' r_{ij}}{\sum_{k \in S} c_k' r_{kj}}. \tag{4.2}$$

This is not possible unless two things can be assumed: First, that a round of migration does not change the deme sizes, so that $c_j = c_j'$ for all j. This this only the case if migration is conservative or, as in the structured Moran model, if the continuous in or out-flux from a deme precisely chancels out with the growth rate of the deme. Second, that the c_i' in (4.2) refer to actual deme sizes and not effective deme sizes. Hence, the fact that the deme sizes obtained from data are effective deme sizes, that contain all the un-separable factors that determine drift in addition to deme size, makes it impossible to convert a backward matrix to a forward matrix, unless it can be assumed that each deme is panmictic (In which case $c_i N_e = c_i N_T$).

In conclusion, the only the genetic effect of migration, given by the backwards migration matrix, can be obtained from a data set. Any further interpretation in terms forward migration is highly inadvisable.

4.3 General Problems in Retrospective Population Genetic Analysis.

The problems concerning the resolution of information that may be obtained from a data set, described in the beginning of the chapter, is a type of problems that can not be circumvented. A second type of problems in retrospective population genetics are the problems pertaining to the validity of the null-hypothesis, that inferences are based on. If we are to make inferences on the population structure, we have to be able to assume that the effects seen in data owe to structuring and not to other effects. In other words, we must ensure ourselves to the extent possible, whether: (I) the size of the population or and relative sizes of the demes it may be subdivided into have not changed in the evolutionary time perspective of the Coalescent, (II) the backwards migration matrix of a possible structure in the population has not changed in this time perspective either, (III) the sequences considered have not been subject to effects of selection in the time perspective of the Coalescent, and (IV) the sequences are not subject to recombination.

These premises are difficult to establish, and ecological observations are of little help since these can only describe features of the present population, and not the past under study in a Coalescent framework. If such a null-hypothesis can not be established with at least some degree of certainty, inferences from data are of little value, since the population features listed above may produce effects on the Coalescent that obliterate inferences on structure.

Below, some effects that may cause erroneous inferences, if not taken into account, are described in brief.

4.3.1 Non-Constant Deme Sizes and Backwards Migration Matrix

If we assume that the backwards migration matrix is constant through time, we must also assume that the forward migration matrix and the relative sizes of demes that determine the backwards migration matrix are constant in time, see (4.1). However, changing species composition and variations due to disasters or climatic fluctuations, may change the quality of the demes, and along with this, the demographic regimes in the demes. It is not unreasonable to expect that the forward migration matrix will change in accordance with demography, and if this is the case so will the backward migration matrix.

If the quality of a deme may vary over time, so will the number of individuals it may sustain. Hence, for the same reasons as for the forward migration rate, deme sizes may also change through time. These changes in population and deme sizes may not be possible to detect.

If the forward migration matrix or the relative sizes of demes change drastically through time, inferences on structure is not possible. However it may be assumed that these stay the same while only the total population size change. This may be the case if an area has been colonised or exposed to a disaster diminishing the deme size so that the sizes of all demes grow exponentially during the period of time that the sample find its common ancestor. As the deme size decrease backwards in time, the coalescence rate will increase along with it. In a panmictic population this will result in shorter trees with long terminal branches, and a high intensity of coalescences before the root of the tree (Slatkin & Hudson 1991).

In structured populations this effect may confuse inferences on structure, since the prolonging effect of coalescence preclusion on the last branch levels may to some extent cancel out with the shortening effect of exponential growth.

4.3.2 Recombination

If intra-genic recombination occurs, different parts of a sequence will have different genealogical histories. Each genealogy represents a realisation of a stochastic process and is associated with a large variance. Hence, recombining sequences will yield parameter estimates with a smaller variance than non-recombining sequences.

The problems arise when sequences that are assumed not to recombine actually do so. neglecting the effect of recombination will in this case produce trees that superficially resemble those for exponential growth (Schierup & Hein 2000). These trees with long terminal branches will result because shuffling parts of the sequences will make the distances between sequences more alike, resulting in a more star-like genealogy. The two forces may be distinguished by Tajima's D comparing the number of pairwise differences to the number of segregating sites:

$$D = \frac{\Pi - S/a_n}{\sqrt{Var(\Pi - S/a_n)}}, \tag{4.3}$$

where $a_n = \sum_{i=1}^{n}(1/i)$. The mean of D is independent of recombination but will give negative values in the case of exponential growth.

The implications for a structured population are difficult to imagine. However, the effects of recombination must be expected to obscure the effects of structure.

4.3.3 Migration and Historical Association

If a population at some time in the past was divided into two demes, it is difficult to distinguish two situations: First, a situation where the two demes diverged a long time ago but where migration have occured between them since then, and second, a situation where the two demes diverged recently but have been virtually isolated from each other since then. For a range of relocation rates, and divergence times, the mean number of pairwise differences, both for sampling in one deme and in two demes, is the same for the two settings. However, the variances of pairwise differences show a somewhat different dependence on relocation probability and divergence time This implies that sets of relocation probabilities and divergence times that produce the same mean pairwise differences may be separated by variances of pairwise differences (Wakeley 1996).

4.4 Conclusion

An introduction to the Coalescent and the structured Coalescent have been given. Further, the effects of a source-sink functionality on the Coalescent has been described. This was done in a Moran model by expressing all transition probabilities of the structured Coalescent, in terms of the birth and death rates given by the demographic regimes in each deme. The effects of a source-sink functionality on tree structure in small demes is described, and a result for the source-sink effective population size has been given.

The Coalescent is a powerful tool in population genetics. It is simple, and describes the ancestral relationship of the sampled sequences. It must, however, be used with caution. As discussed above, it is difficult to establish whether the assumptions, that inference on the sequence sample is based on, actually hold. Effects such as non-constant migration regimes, non-constant deme sizes, historical association, recombination and selection, will obscure the information on structure in the sample, if not taken into account. Hence, even though structured populations are best described by the structured Coalescent, additional forces such as the above may obliterate the possibility of inference.

Bibliography

Donnelly, P. & Tavaré, S. (1995), 'Coalescents and the genelogical structure under neutrality', *Annu. Rev. Genet.* **29**, 401–421.

Fu, Y.-X. & Li, W.-H. (1993), 'Statistical tests of neutrality of mutations', *Genetics* **133**, 693–709.

Hey, J. (1991), 'A multi-dimentional process applied to multi-allelic selection models and migration models', *Theor. Popul. Biol.* **39**, 30–48.

Hudson, R. R. (1990), Gene genealogies and the coalescent, *in* D. J. Futuyma & J. Antonovics, eds, 'Oxford Surveys in Evolutionary Biology', Vol. 7, Oxford Univ. Press, Oxford.

Kimura, M. (1969), 'The number of heterozygous nucleotide sites in a finite population due to steady flux of mutations', *Genetics* **61**, 893–903.

Kingman, J. F. C. (1982*a*), 'The coalescent', *Stocastic Prosess. Appl.* **13**, 235–248.

Kingman, J. F. C. (1982*b*), 'On the genealogy of large populations', *J. Appl. Prob.* **19A**, 27–43.

Li, W.-H. (1976), 'Distribution of nucleotide differences between two randomly chosen cistrons in a subdivided population: The finite island model', *Theor. Popul. Biol.* **10**, 303–308.

Li, W.-H. (1997), *Molecular Evolution*, 1. edn, Sinauer Associates, Sunderland.

Nagylaki, T. (1980), 'The strong migration limit in geographically structured populations', *J. Math. Biol.* **9**, 101–114.

Nagylaki, T. (1998), 'The expected number of heterozygous sites in a subdivided population', *Genetics* **149(3)**, 1599–1604.

Nordborg, M. (1997), 'Structured coalescent processes on different time scales', *Genetics* **146**, 1501–1514.

Notohara, M. (1990), 'The coalescent and the genealogical process in geograhpically structured populations', *J. Math. Biol.* **29**, 59–75.

Notohara, M. (1993), 'The strong-migration limit for the genealogical process in geographically structured populations', *J. Math. Biol.* **31**, 115–122.

Pulliam, R. H. (1988), 'Sources, sinks, and population regulation', *American Naturalist* **132**, 652–661.

Schierup, M. & Hein, J. (2000), 'Consequences of recombination on traditional phylogenetic analysis', *Genetics* **156**, 879–891.

Slatkin, M. (1987), 'The average number of sites seperating DNA sequences drawn from a subdivided population', *Theor. Popul. Biol.* **32**, 42–49.

Slatkin, M. (1989), 'Detecting small amounts of gene flow from phylogenies of alleles', *Genetics* **121**, 609–612.

Slatkin, M. (1991), 'Inbreeding coefficients and coalescence times', *Genet. Res. Camb.* **58**, 167–175.

Slatkin, M. & Hudson, R. R. (1991), 'Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations', *Genetics* **129**, 555–562.

Slatkin, M. & Maddison, W. P. (1989), 'A cladistic measure of gene flow inferred from the phylogenies of alleles', *Genetics* **123**, 603–613.

Strobeck, C. (1987), 'Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision', *Genetics* **117**, 149–153.

Tajima, F. (1983), 'Evolutionary relationship of DNA sequences in finite populations', *Genetics* **105**, 437–460.

Tajima, F. (1989), 'DNA polymorphism in a subdivided population: The expected number of segregating sites in the two-subpopulation model', *Genetics* **123**, 229–240.

Takahata, N. (1988), 'The coalescent in two partially isolated diffution populations', *Genet. Res. Camb.* **52**, 213–222.

Takahata, N. (1991), 'Genealogy of neutral genes and spreading of selected mututations in a geographically structured population', *Genetics* **129**, 585–595.

Tavaré, S. (1988), 'Line-of-decent and genealogical processes, and their applications in populations genetics models', *Theor. Popul. Biol.* **26**, 119–164.

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

Wakeley, J. (1998), 'Segregating sites in wright's island model', *Theor. Popul. Biol.* **53**, 166–174.

Wakeley, J. (1999), ‘Nonequilibrium migration in human history’, *Genetics* **153**, 1863–1871.

Wakeley, J. (2001), ‘The coalescent in an island model of population subdivision with variation among demes’, *Theor. Popul. Biol.* **59**, 133–144.

Watterson, G. A. (1975), ‘On the number of segregating sites in genetical models without recombination’, *Theor. Popul. Biol.* **7**, 256–276.

Wilkinson-Herbots, H. M. (1998), ‘Genealogy and subpopulation differentiation under various models of population structure’, *J. Math. Biol.* **37**(6), 535–585.