Mathematical Population Genetics II

Lecture Notes

Joachim Hermisson

June 9, 2018

University of Vienna
Mathematics Department
Oskar-Morgenstern-Platz 1
1090 Vienna, Austria

These lecture notes are mainly based on previous course material from the "Population Genetics Tutorial" by Peter Pfaffelhuber, Pleuni Pennings, and JH (2009), from the lecture "Migration-Selection Models in Population Genetics" by Reinhard Bürger (2010), and from the book "Probability Models for DNA Sequence Evolution" by Rick Durrett (2008).

1 Structured populations

Individuals in natural populations differ from one another in very many ways. While some of these differences are irrelevant for the future fate of the population, others are crucial. It is one of the main challenges of ecological and evolutionary theory alike to identify the most relevant differences and to include them into a model. In evolutionary models, the main focus is usually on differences among individual genotypes. After all, the study of changes in the genetic composition of a population is the objective of population genetics. Accordingly, a major part of population genetic theory ignores all other differences among individuals. Mathematically, this means that individuals with the same genotype are *exchangeable*, i.e., the model (and all results that can be derived from the model) are invariant under arbitrary permutations of these individuals. Obviously, this assumption leads to a significant simplification of the mathematical formalism. For many or most natural populations, however, this is a gross over-simplification.

In the first part of this lecture, we will take a closer look at the impact of non-genetic differences among individuals on the course of evolution. Maybe the most important aspect of population structure is space: individuals live in certain local neighborhoods, and they are more likely to mate or to compete with conspecifics that live in this same neighborhood than with individuals that live in remote parts of the habitat. Most of our models below will thus treat spatial structure, but we will also see that ploidy levels and the mode of reproduction (inbreeding or outbreeding) can lead to similar effects. Other aspects of structure include the age distribution of a population. Our main goal in the first part is to describe neutral evolution in a structured population and we will mainly use stochastic theory of the structured coalescent to address these issue.

In a second part of the lecture, we will turn to a set of problems that is at first sight quite unrelated: complex multi-locus genotypes under selection. We will see that methods for population structure can also be applied to this case. The crucial key idea is to split the a genotype into a focal gene and its genetic background. The genetic background then takes the role of the "habitat" an allele at the focal gene lives in. In the course of evolution, alleles can switch backgrounds by recombination. ths leads us to models of background selectio and selective sweeps that are used to describe the impact of selection on patterns of nucleotide diversity in populations and to infer event of past selection from these patterns. Finally, in a third part of the lecture we provide a brief introduction to models on selection in a spatially structured population.

1.1 No structure

Before talking about the effects of population structure, we should first clarify what the absence of structure means. Mathematically, a population model has no structure if all individuals with the same genotype are exchangeable. We can then subsume all these individuals in a single number: the frequency of a genotype. Biologically, this implies that competition and matings among individuals should only depend on genetic factors – and random chance.

In the ideal diploid Wright-Fisher model mating is random and a new diploid individual is formed by combining two random haploid gametes (taken from diploid parents). If two alleles A and a at a locus occur with frequencies p and $q = 1 - p$ we obtain the following expected frequencies for the genotypes in the offspring generation:

$$\begin{aligned} p^2 &\quad \text{for genotype} \quad AA, \\ 2pq &\quad \text{for genotype} \quad Aa, \\ q^2 &\quad \text{for genotype} \quad aa. \end{aligned}$$

Random mating (or *panmixia*), in particular, is a mating scheme that guarantees absence of population structure for this phase of the life cycle. In terms of the coalescent, this assumption means that any two alleles from the offspring generation have the same probability to have a common ancestor in the previous generation. Forward in time, a single generation of random mating in a diploid population (two generations with separate sexes) leads to Hardy-Weinberg proportions in the expected genotype frequencies of the offspring generation (before selection). Vice-versa, approximate Hardy-Weinberg (HW) proportions in a natural population are often taken as evidence of absence of population structure. Due to genetic drift, the match will never be perfect in a finite population (or any population sample), but a standard χ^2 test easily answers the question whether the differences are significant. Deviations from HW proportions can have causes other than population structure, such as assortative mating based on genotypes, or other types of selection on mate choice or fertility. However, since HW equilibrium is restored in a single generation,only factors that affect the population *in the present generation* will matter and effects do not easily add up across generations. This is in contrast to deviation from linkage equilibrium, which has a much longer memory (and therefore can tell us more about the history of a population).

Global competition means that all individuals (with the same genotype) are equal and exchangeable with respect to selection. I.e., selection during a single generation does not depend on any non-genetic variables, such as space. If selection acts at a different life stage than mating, we can have global competition either with random mating or with non-random mating. An example for the latter are species like salmon, where individuals mate at the place where they were born, but mix and compete during other phases of their life cycle.

1.2 A general model of population structure

Imagine a population of diploid size N, corresponding to $2N$ gene copies (haploids). Evolution proceeds in discrete generations like in a Wright-Fisher model. To introduce population structure, we assume that individuals carry non-genetic labels. These group labels can affect the life at various stages. With local competition, only (or primarily) individuals in the same group compete for a common resource. With group-structured mating, individuals with the same group label are more (or also less) likely to mate. In general, the life-cycle could be modeled as follows:

1. We census our population in the zygote state of newborn offspring. All individuals belong to one or several groups, indicated by a group label or parameter.

2. Each group produces an infinite offspring pool (gametes) and offspring inherit the group label. Back in time, only offspring with the same group label can coalesce.

3. Selection and mutation (if included in the model) happen on the level of the infinite offspring pools. Both change the frequencies of the allelic composition in the pools. Selection either happens during the gametic life stage or already during reproduction via differential adult fertilities. Importantly, selection can be local and act only among individuals of the same group. Alternatively, selection can be global and change the size (or weight) of different offspring pools relative to each other.

4. Finally, individuals (zygotes) in the offspring generation are sampled from the infinite offspring pools. Offspring individuals are assigned to new groups. The sampling rules define whether and how alleles from different offspring pools can mix and enter a common group in the new generation. This is, they define how much gene-flow (genetic exchange) among groups will occur.

This framework comprises a large number of specific models, in particular:

- **Fixed deme structure.** In the standard model for spatial structure, the population inhabits a finite number of demes (or islands) of fixed size. Both mating and competition occur locally on the island, genetic exchange is possible through migration. We will discuss this model in detail in section 1.4.

- **Levene model: No structure during reproduction.** If we assume that assignment to new groups is independent of the individual labels, there is no population structure with respect to mating. This is a limiting case of the fixed deme model for strong migration. The group labels are only relevant for the selection phase, which occurs within each deme (soft selection). After selection, they become irrelevant: We can join all gamete pools and sample from the joint pool to obtain a new generation of zygotes.

- **Hard selection: No structure during competition.** In the opposite extreme, we assume that there is global competition for resources among individuals of all groups. This means that the group labels do not influence the offspring distribution of any individual: They do not affect decent. We thus obtain the same distribution of coalescent histories as without labels if we start with a random sample from the total population. However, the labels are informative of relatedness (they *reflect* descent). Usually, individuals with the same label will be closer related, on average, than random individuals. We thus get changed coalescent histories if we do not sample randomly with respect to labels.

- **Selfing model: Diploids as groups.** Here, each diploid individual is treated as a group of size 2. Each diploid produces its own offspring pool. For the next generation, diploid offspring are produced by sampling pairs of individuals from the old pools. Both come from the same offspring pool with the selfing probability p_s. We will further expand this approach in the next section.

Our general model shows that population structure can either act together with selection, but also entirely on its own, to affect the genealogical relationships in a population. In the following sections, we will first discuss in some detail how population structure acts under neutrality. Only in part 3 of the lecture, we will also discuss models that combine population structure and selection. An important aim of population genetic theory is to link an evolutionary model (which may include population structure) to observable patterns in DNA diversity. For a neutral model, this can be done in two steps: in the first step, we ask how population structure influences the genealogical relationships among individuals (*decent*). These relationships exist independently of allelic types (*state*). Allelic types (mutations) are only added in the second step. As we will see below, the skews in genealogies due to population structure often lead to characteristic patterns in polymorphism and diversity. It is important, however, to recognize these patterns as a consequence of structure, rather than its definition.

1.3 Inbreeding

The mating of relatives to produce offspring is referred to as *inbreeding.* The most extreme form of inbreeding results from self-fertilization which is possible in many plants but also in snails and fungi. But also any other scenario that leads to higher relatedness of mated pairs than expected for random picks from the population will induce inbreeding. There are several such scenarios, with should be distinguished due to their diverging implications on population structure. Note first that spatial population structure with local matings and limited dispersal of offspring will lead to inbreeding as a by-product. Alternatively, inbreeding can be the primary effect if preferred matings among relatives (or avoidance thereof) is a property of the mating system. Since this makes matings dependent on a non-genetic factor (because, e.g., relatives are preferred over strangers even if the latter are genetically identical), it induces population structure – but no spatial structure if dispersal of offspring is global. Finally, assortative mating based on similarities in genotype or phenotype leads to inbreeding without any population structure according to our definition above.

Let us focus on the second scenario, where inbreeding is the primary effect rather than a by-product. Consider a population of diploids. Two individuals are related of degree k if they have a (most recent) common ancestor k generations ago. Assume that there is non-random mating of related individuals up to a certain degree of relatedness (such a full/half sibs, cousins, etc), but random mating among more distantly related individuals. To quantify the strength of inbreeding, we define the *inbreeding coefficient f* as the probability that two homologous alleles in an individual are identical by descent *due to inbreeding.* To formalize this, assign labels to family groups, where two alleles belong to the same family if and only if the individuals they belong to are related up to the degree considered. Note that groups are not necessarily disjoint: an allele can belong to multiple family groups. Now follow the ancestry of the two homologous alleles in the focal individual back in time as long as the both lines still belong to a common family group. Define f as the probability that both lines coalesce during this time.

There are two main ways how this definition can be used:

1. If the pedigree of the focal individual is known up to some founder generation, where all ancestors are (assumed to be) unrelated, we can calculate the *individual inbreeding coefficient* $f = f_i$, which tells us the average level of inbreeding along the genome of this focal individual. This is usually done in breeding programs where pedigrees are recorded and the founder generation may represent the start of the program from a sample of wild-caught animals or collected plants. Depending on the pedigree, the inbreeding coefficient of different individuals will usually be different.

2. Alternatively, if we know about average mating probabilities among relatives in a wild (or managed) population, we can derive a population-wide average inbreeding coefficient, which we will simply denote as f. This is what is usually done in population genetic models of natural populations.

3. One can easily extend the concept of the inbreeding coefficient by considering excess identity by descent among two homologous alleles in related individuals (such as in sibs) rather than in the same individual. The coefficient is also called *coancestry coefficient* in this context.

Depending on the mating system, the derivation of the inbreeding coefficient can lead to lengthly calculations. We will only consider two simple cases, self-fertilization and full-sib mating, where the analysis is simple.

Selfing Assume that in a diploid population fertilization can occur by either random mating or selfing, and that selfing occurs with probability p_s. Alleles then form “family” groups of two, which sit in the same individual. With probability p_s, mating occurs within the group, and with probability $1 - p_s$ it occurs with an allele from a different group. Backward in time, the probability that two homologous alleles in the same offspring individual derive from the same allele in the parent generation is $p_s/2$. Indeed, we need, first, that both alleles are from the same diploid parent (probability p_s), and, second, that both are copies from the same parental allele (probability 1/2 with random segregation). The probability that both lines of descent have not yet coalesced, but are still in the same group (= same individual) is also $p_s/2$. We thus find

$$f = \frac{p_s}{2} + \left(\frac{p_s}{2}\right)^2 + \left(\frac{p_s}{2}\right)^3 + \cdots = \frac{p_s}{2 - p_s}\,. \tag{1.1}$$

As expected, the inbreeding coefficient can vary between $f = 0$ for obligate outcrossers, $p_s = 0$, and $f = 1$ for strict selfers, $p_s = 1$. Note that random mating (random union of gametes), as assumed in the diploid Wright-Fisher model, corresponds to $p_s = 1/N$ and thus a slightly positive value for f.

- An even simpler model of selfing results if we assume that all sperms and eggs of an individual are genetically identical, e.g., because they are produced during a haploid stage of the life cycle. If p_s is the probability that an egg is fertilized by a sperm from the same individual (selfing rate), we simply have $f = p_s$ in this case. For convenience, this scheme is often used in population genetic models to account for effects of inbreeding.

Full-sib mating Consider again a diploid sexual population. Selfing is not possible, but there is a probability p_1 that mating occurs among full sibs (first-degree relatives). Backward in time, two homologous alleles in the same individual will thus go back to full sibs in the previous generation with probability p_i. Consider now two homologous alleles from full sibs: With probability 1/4, they will coalesce in the previous generation. With probability 1/4 they will go back to the same individual, but not coalesce. Finally, with probability $p_i/2$ they will go back to full sibs again. Call the coancestry coefficient for full sibs f_{fs}. We then have

$$f = p_1 f_{\mathrm{fs}}\,, \tag{1.2}$$

$$f_{\mathrm{fs}} = \frac{1}{4} + \frac{1}{4}\,f + \frac{p_1}{2}\,f_{\mathrm{fs}}\,, \tag{1.3}$$

which derives to

$$f = p_1 f_{\mathrm{fs}} = \frac{p_1}{4 - 3p_1}\,. \tag{1.4}$$

Inbreeding and the coalescent

The effect of (family) group structure and inbreding on the coalescent becomes apparent if we compare the genealogies of pairs of homologous alleles that are either taken from a single diploid individual or randomly drawn from the entire population. Let $E[T_I]$ and $E[T_T]$ be the expected coalescent times for these two cases, respectively. We consider the case of two homologs from a single diploid first. Back in time, one of two events will occur: with probability f both lines of descent will coalesce within the group of related ancestors, while with probability $(1 - f)$ both lines will enter unrelated ancestors at some point in time. Let T_0 be the time in generations when either of these events happens. We then have

$$E[T_I] = E[T_0] + (1 - f)E[T_T] \tag{1.5}$$

(where we ignore the case that a random pick from the entire population may come from related individuals). We can now define

$$f_{IT} = \frac{E[T_T] - E[T_I]}{E[T_T]} = f - \frac{E[T_0]}{E[T_T]} \approx f\,. \tag{1.6}$$

f thus measures (approximately) the percentage reduction in the expected coalescence time of two genes in the same individual due to inbreeding. The approximation is based on the fact the we generally have $E[T_0] \ll E[T_T]$. Exact results can be calculated for specific models. E.g., for the diploid selfing model we have

$$E[T_0] = \frac{1}{(1-p_s)+p_s/2} = \frac{2}{2-p_s} = 1+f \tag{1.7}$$

$$E[T_T] = N + \frac{1}{2}\Big(E[T_0] + (1-f)\,E[T_T]\Big) = \frac{2N}{1+f} + 1 \approx \frac{2N}{1+f}\,. \tag{1.8}$$

($E[T_T]$ reduces to $2N$ with $f = 1/(2N-1)$ as it should). Thus, the term $E[T_0]/E[T_T] = (1+f)^2/(2N+1+f)$ only contributes a correction of order $1/N$. If we measure time in units of the haploid population size, $\tau = t/(2N)$, and apply the usual coalescent limit $N \to \infty$, the time $E[T_0]$ scales to zero and can be ignored. This holds more generally for models of inbreeding also beyond selfing. For the coalescent, we then obtain a combined process on two time scales:

- Coalescent events among genes in related individuals due to inbreeding occur instantaneously.
- Coalescent times among genes in unrelated individuals, and among related genes that do not coalesce due to inbreeding, are rescaled by a factor that depends on the inbreeding coefficient. For selfing, in particular, we see that coalescent times of such genes are reduced by $(1+f)$.

If we want to construct the full coalescent process for a larger sample from a partially selfing population, we need to distinguish two scenarios, depending on the sampling procedure.

1. If all gene copies in the sample are taken from different individuals, all coalescent event happen on the "slow" time scale. Inbreeding than affects the coalescent process only by the rescaling of the coalescent times by a unique factor. Alternatively, we can capture this time rescaling via the definition of the *coalescent effective population size* N_e^c, which is the population size of a standard neutral Wright-Fisher model that leads to the same coalescent. For partial selfing, in particular, we obtain

$$N_e^{(c)} = \frac{N}{1+f}\,.$$

   Since $N_e^{(c)} < N$ for $f > 0$, coalescence in partially selfing populations is faster since pairs of genes only need to trace back to a common *diploid* ancestor and then have a chance of $(1+f)/2 > 1/2$ to coalesce instantaneously (i.e., within time T_0). Selfing thus leads to smaller coalescence trees and thus to reduced polymorphism, but not to any deviations in the site-frequency spectrum relative to a standard random mating population.

2. A different situation arises if the sample from a partially selfing population contains both allele copies from (some or all of) the diploid individuals. We then need to take the fast time scale into account, which may lead to the immediate coalescence of some the related genes. Since coalescence due to selfing occurs with probability f per diploid, the number of these coalescent events in a sample of n diploids is binomially distributed with parameters f and n. The probability that we have k fast coalescent events is thus

   $$\Pr[k] = \binom{n}{k} f^k (1-f)^{n-k} .$$

   If we ignore mutation on the fast time scale, each of these fast coalescent events leads to a genotype that is represented twice in the sample. After the fast initial phase, we thus have $2n-k$ lines of descent, of which k have two descendants and $2(n-k)$ a single descendant. Since all these $2n-k$ lines sit in different individuals, they now enter the coalescent process on the slow time scale as described above. In this second phase, the haplotypes are thus once again connected by the standard neutral Kingman coalescent with effective size $N_e^{(c)} = N/(1+f)$. Site frequency statistics of the combined process running through both phases can easily be obtained by combining the binomial sampling step with the statistics of the standard neutral spectrum.

The coalescent for other types of inbreeding, such as partial sib mating, can be dealt with in an analogous way. In more general, the separation of time scales with a fast phase to describe coalescence processes within groups and a slow phase to describe coalescence events among genes from different groups is a typical feature of structured coalescent events. We will meet further examples in the following sections.

1.4 Spatial population structure

Spatial population structure is an ubiquitous property of populations that live in extended areas. There are two main ways to incorporate spatial structure into a population genetic model. Deme structured models assume that the total metapopulation can be divided into discrete panmictic subpopulations that are connected by limited migration. Alternatively, models in continuous space and time describe populations in a diffusion setting. The crucial property of spatial models is that selection and/or mating occurs primarily among individuals in the same spatial neighborhood. As in the case of inbreeding, we can define a fixation index to measure the consequences of spatial structure on diversity patterns. In this lecture, we will focus on discrete deme models. The crucial advantage of deme models is that they allow for a study of population genealogies within the coalescent framework. Although there is some recent progress to set up a coalescent theory in continuous space (by Etheridge, Barton and coworkers) this is exceedingly complicated.

Spatial structure and local competition

We will now define our standard model to describe spatial population structure. Assume that a monoecious, diploid population of size N is structured into d demes (patches, colonies, islands) of constant size $N_1, \dots, N_d$, with $\sum_i N_i = N$. Generations are discrete and the life cycle corresponds to the one of the structured Wright-Fisher model. We will focus on neutral evolution, and, for the time being, will also ignore different allelic states altogether. Dispersal is defined via the so-called backward migration matrix. For this, let $m_{ij} \geq 0$ designate the probability that a gamete in deme i after dispersion was produced by an adult from deme j and $m_{ii} = 1 - \sum_{j \neq i} m_{ij} \leq 1$ (the m_{ij} are also called *backward migration fractions*). Then

$$\mathbf{M} = (m_{ij})$$

is a stochastic matrix. It thus has $\lambda_{\max} = 1$ as largest eigenvalue with right eigenvector $(1, 1, \dots, 1)^T$. We will usually assume that $\mathbf{M}$ is ergodic, i.e., irreducible and aperiodic. Irreducibility means that descendants from every individual are eventually able to reach any other deme. Aperiodicity means that there are no periodic cycles. Aperiodicity is already guaranteed if $m_{ii} > 0$ for at least one i, which is biologically trivial. The Perron-Frobenius theorem then implies that $\lambda_{\max} = 1$ is simple and all other eigenvalues are smaller in absolute value. Furthermore, $\mathbf{M}$ has a positive left eigenvector $\mathbf{u}$ for $\lambda_{\max}$, which corresponds to the stationary distribution of the backward migration process. We normalize $\mathbf{u}$ as a probability vector, $\sum_i u_i = 1$. We assume that dispersal is followed by random union of gametes in each deme.

- Backward migration is connected to forward migration rates $q_{j \to i}$ (the probability that an offspring of an adult individual from deme j migrates to deme i) as

  $$m_{ij} = \frac{N_j q_{j \to i}}{\sum_k N_k q_{k \to i}}\,. \tag{1.9}$$

  Note that migration as defined here does not affect the (fixed) population sizes of the demes. Forward migration of an offspring does not imply that the migrant will find a spot in its target deme. The ratio accounts for sampling from the migrant pool.

- Let $c_i = N_i/N$ be the relative deme sizes and $\mathbf{c} = (c_1, \dots, c_d)$ the corresponding row-vector. Then the components of $\mathbf{c}^{(t)} = \mathbf{c}\mathbf{M}^t$, $c_i^{(t)}$, give the expected contributions of deme-i ancestors t generations ago to the current population. If $\mathbf{M}$ is ergodic, $\mathbf{c}^{(t)}$ converges to the stationary distribution $\mathbf{u}$ as $t \to \infty$. Demes contribute to this distribution according to their size if and only if $\mathbf{u} = \mathbf{c}$. In this case, we have $\sum_{k \neq j} N_k m_{kj} = N_j(1 - m_{jj})$; in words: the expected number of migrants with parent in deme j equals the expected number of immigrants into deme j. This is also called *conservative migration*. An alternative interpretation of u_i is the proportion of time that a line of descent of any current individual will spend in deme i.

- The stationary distribution depends only on the relative migration rates. It thus remains unchanged if we rescale all m_{ij} with $i \neq j$ to αm_{ij} for some $\alpha > 0$. To see this, note that the rescaled migration matrix can be written as $\tilde{\mathbf{M}} = \alpha(\mathbf{M} - \mathbf{1}) + \mathbf{1}$, where $\mathbf{1}$ is the identity matrix. $\tilde{\mathbf{M}}$ and $\mathbf{M}$ have the same eigenvectors with a transformed spectrum $\tilde{\lambda} = \alpha(\lambda - 1) + 1$. In particular, $\tilde{\lambda}_{\max} = \lambda_{\max} = 1$. Rescaling by α only changes the time-scale on which the stationary distribution is approached.

- We can calculate the single-generation coalescence probability for two randomly chosen alleles as
$$p_{c,1} = \sum_i (c_i^{(1)})^2 \frac{1}{2N_i}$$
and the inbreeding effective population size as the inverse $N_e^{(i)} = 1/(2p_{c,1})$. For conservative migration with $\mathbf{u} = \mathbf{c}^{(1)} = \mathbf{c}$, we obtain $p_{c,1} = 1/(2N)$ and thus $N_e^{(i)} = N$. This shows that population structure (even extremely strong one with vanishing migration) can go entirely unnoticed by measures like $N_e^{(i)}$.

Similarly to the case of inbreeding, we can define a probability f_i that two alleles taken from deme i coalesce within this deme before emigrating backward in time. We obtain

$$f_i = \frac{m_{ii}^2/(2N_i)}{(1 - m_{ii}^2) + m_{ii}^2/(2N_i)}\,. \tag{1.10}$$

Coalescent time scale. To simplify the resulting expressions and to make further progress, we switch to a continuous-time process on the so-called coalescent time scale. Let t be time measured in generations and

$$t = \lfloor 2N\tau \rfloor\,, \tag{1.11}$$

where τ is a continuous time parameter, N is population size, and $\lfloor x \rfloor$ denotes the largest integer smaller than $x \in \mathbb{R}$. Let p_0 be the probability (per generation) of a focal event of interest in the genealogy and define a rescaled quantity $P_0 = 2Np_0$. Then the time t_0 to the focal event is geometrically distributed

$$\Pr[t_0 > t] = \Big(1 - p_0\Big)^t = \Big(1 - \frac{P_0}{2N}\Big)^{\lfloor 2N\tau \rfloor} = \exp\big[-P_0\tau\big] + \mathcal{O}[N^{-1}]\,.$$

We can therefore define an exponentially distributed continuous random variable T_0 for the time to the focal event on the τ-scale,

$$\Pr[T_0 > \tau] = \exp\big[-P_0\tau\big]\,,$$

which approximates the original distribution of t_0 up to correction terms of the order of $1/N$. The distribution of T_0 (and the one of t_0 to leading order) does not explicitly depend on the population size, but only on the composite parameter $P_0 = 2Np_o$. Note that P_0 is no longer a probability like p_0, but has the interpretation of a rate.

Applying this scaling procedure to backward migration, we define rescaled migration rates $M_{ij} := 4Nm_{ij}$ $(i \neq j)$. The factor of $4N$ instead of $2N$ is for reasons of consistency with the population genetic literature: M_{ij} thus corresponds to twice the migration rate on the coalescent time scale. The coalescent probability $1/(2N_i)$ in deme i rescales to a rate of $2N/(2N_i) = 1/c_i$. In the continuous time process, events in the genealogy (here: migration and coalescence) never happen at the same time. We can therefore define a combined event (such as coalescence *or* migration) and obtain its rate simply by adding the rates of the single events. Starting with a sample of size 2 in deme i, the total rate of events is given by $(1/c_i) + \sum_{j\neq i} M_{ij}$. The expected time to the first event (coalescence or emigration) is given by the inverse of the total rate,

$$\mathrm{E}[T_{i,0}] = \frac{1}{(1/c_i) + \sum_{j\neq i} M_{ij}} \,. \tag{1.12}$$

The probability that this event is coalescence rather then migration derives in continuous time as the ratio of the coalescence rate to the total rate, thus

$$f_i = \frac{1/c_i}{(1/c_i) + \sum_{j\neq i} M_{ij}} = \frac{1}{c_i \sum_{j\neq i} M_{ij} + 1} \,, \tag{1.13}$$

which corresponds to the discrete time expression (1.10) up to correction terms of order N^{-1}.

Pairwise coalescence times. For a general deme-structured model, we define T_{ij} as the coalescence time (on the coalescence time scale) for two random alleles taken from deme i and j, respectively. We then find the following recursions for the expected values of the T_{ij},

$$\mathrm{E}[T_{ii}] = \mathrm{E}[T_{i,0}] + \frac{1-f_i}{\sum_{j\neq i} M_{ij}} \sum_{j\neq i} M_{ij}\mathrm{E}[T_{ij}] \,, \tag{1.14}$$

$$\mathrm{E}[T_{ij}] = \frac{2}{\sum_{k\neq i} M_{ik} + \sum_{k\neq j} M_{jk}} + \frac{\sum_{k\neq i} M_{ik}\mathrm{E}[T_{kj}] + \sum_{k\neq j} M_{jk}\mathrm{E}[T_{ik}]}{\sum_{k\neq i} M_{ik} + \sum_{k\neq j} M_{jk}} \quad \text{for } i \neq j \,. \tag{1.15}$$

In general, this is a linear equation system of order d^2. It can only be solved for some special cases. We can further define coalescence times for a random pair of individuals taken either from the same deme (T_S) or from the total population (T_T) as

$$T_S = \sum_i c_i T_{ii} \quad ; \quad T_T = \sum_{i,j} c_i c_j T_{ij} \,. \tag{1.16}$$

A measure for the strength of population structure is the given by

$$f_{ST} = \frac{\mathrm{E}[T_T] - \mathrm{E}[T_S]}{\mathrm{E}[T_T]} \,, \tag{1.17}$$

which compares the coalescent times for pairs of genes from the same deme and the total population, respectively. f_{ST} varies from 0 in the absence of structure to 1 for populations with extreme structure, where the ancestries of genes from different demes are completely separated.

Symmetric island model

The symmetric island model makes two crucial assumptions:

1. All demes have equal size, $N_i = N/d$, or $c_i = 1/d$, $\forall i$.
2. Backward migration among all demes is equal, $M_{ij} = M$ for $i \neq j$.

Biologically, we thus assume that there are no close or distant demes (no isolation by distance). Individuals in the offspring generation are either drawn from local parents or they immigrate from a migrant pool that is common to all islands. The total migration rate (forward or backward) is $(d-1)M$. Under these assumptions, the recursions derived above simplify considerably. We only need to distinguish coalescent times within a subpopulation, $T_S = T_{ii}$ and between demes, $T_B = T_{ij}$, $i \neq j$,

$$\mathrm{E}[T_S] = \frac{1}{d+(d-1)M} + (1 - \frac{d}{d+(d-1)M})\mathrm{E}[T_B]\,, \tag{1.18}$$

$$\mathrm{E}[T_B] = \frac{1}{(d-1)M} + \frac{(d-2)M\,\mathrm{E}[T_B] + M\,\mathrm{E}[T_S]}{(d-1)M}\,. \tag{1.19}$$

This system is easily solved to give

$$\mathrm{E}[T_S] = 1 \quad \text{and} \quad \mathrm{E}[T_B] = 1 + \frac{1}{M} \tag{1.20}$$

on the coalescence time scale of $2N$ generations. With $T_T = ((d-1)T_B + T_S)/d$ and $E[T_T] = 1 + (d-1)/(dM)$, we further obtain

$$f_{ST} = \frac{(d-1)/(dM)}{1+(d-1)/(dM)} = \frac{1}{1+Md/(d-1)} \approx \frac{1}{1+M} \tag{1.21}$$

for many demes $d \gg 1$. We have $f_i = d/((d-1)M + d)$ and so again $f_{ST} \approx f_i$, where both are equal in the limit $d \to \infty$ (when the probability for sampling twice from the same deme in T_T becomes negligible). From the measure of f_{ST} we see that population structure is highly relevant whenever $M \lessapprox 1$. On the other hand, it quickly becomes less relevant (with $f_{ST} \to 0$) once $M \gg 1$. Since $M = 4Nm$ on the generation scale, one typically concludes that strong population structure requires less that a single migrant among subpopulations per generation, independently of the population size. Especially for large populations this remarkable. Note, however, that the result only holds under the assumptions of the symmetric island model.

Another remarkable result for the island model is that the expected coalescence time for two individuals from the same deme does not depend on the population structure at all. Two effects of population structure exactly cancel: On the one hand, low migration rates enhance the probability that both lines coalesce locally before one of them emigrates; on the other hand, it may take a long time for both lines to meet again in the same deme once one line has migrated out. Indeed, this property can be proved for an even larger class of models.

Proposition *The result $E[T_S] = 1$, independently of the migration rates, holds in more general for a model with (1) demes of equal size and (2a) symmetric migration $m_{ij} = m_{ji}$ – or even more generally (2b) for every doubly stochastic, ergodic matrix with a uniform stationary distribution of the migration process.*

Proof Consider the Markov process of backward migration only (without coalescence) and follow the lines of descent of two individuals. Since the equilibrium distribution of the backward process is uniform, each line will visit each deme with frequency $1/d$. Since the process is ergodic, also the probability that both ancestral lines are in the same deme is $1/d$. Next, consider an independent Poisson process with rate d ($= 1/(2N)$ in time units of $2N$). Then a fraction of $1/d$ of the Poisson events will fall into time intervals where both lines are in the same deme. Thus, the expected time between those Poisson events *with lines in the same deme* is $d/d = 1$. Because of the Markov property, the same holds true for the expected time to the next such event, given that we both lines are currently in the same (randomly chosen) deme. Since coalescence can be identified with the first Poisson event where both lines are in the same deme, and since T_S averages over demes, we conclude that $\mathrm{E}[T_S] = 1$.

- Note that the proposition does not imply $\mathrm{E}[T_{ii}] = \mathrm{E}[T_{jj}]$ for $i \neq j$. Indeed, this is already violated for three demes in a row with uniform migration between neighboring demes. It follows, however, for all models with equivalent demes.

To derive the variances of T_S and T_B, we can again use the independence of the times to consecutive events in the Markov process,

$$\begin{aligned}
\mathrm{Var}[T_S] &= \mathrm{Var}[T_{i,0}] + (1-f)\mathrm{E}[T_B^2] + f\cdot 0^2 - ((1-f)\mathrm{E}[T_B] + f\cdot 0)^2 \\
&= \Big(\frac{1}{d+(d-1)M}\Big)^2 + \frac{(d-1)M}{d+(d-1)M}\Big(\mathrm{Var}[T_B] + \frac{d(1+1/M)^2}{d+(d-1)M}\Big) \\
&= \frac{d(d-1)(M+2+1/M)+1}{(d+(d-1)M)^2} + \frac{(d-1)M}{d+(d-1)M}\mathrm{Var}[T_B]\,, \qquad (1.22)\\
\mathrm{Var}[T_B] &= \frac{1}{M^2} + \mathrm{Var}[T_S]\,. \qquad (1.23)
\end{aligned}$$

This results in

$$\mathrm{Var}[T_S] = 1 + \frac{2(d-1)}{dM}\,, \qquad (1.24)$$

$$\mathrm{Var}[T_B] = \frac{1}{M^2} + 1 + \frac{2(d-1)}{dM}\,. \qquad (1.25)$$

The variance of T_S is thus not independent of the population structure, but increases $\sim (1/M)$. This is expected: for very weak migration, both lines likely coalesce very quickly in the same deme before a migration event happens. However, if migration happens first, the expected coalescence time will be very long.

1-dim stepping stone model

In order to introduce isolation by distance, we need a model with inhomogeneous migration: migration between close patches will be higher than migration between distant patches. The easiest way do this is to place demes of equal size on a regular grid or lattice and only allow migration between nearest-neighbor demes. This is the so-called stepping-stone model, first introduced by Kimura (1953). The easiest grid that can be imagined is just a one-dimensional chain.

Suppose we have d demes, with N/d individuals each, connected in a chain. Migration only occurs between neighboring demes, with (forward or backward) rate M in both directions. We still need to decide on the migration pattern for the boundary demes. The most convenient choice are periodic boundary conditions, i.e., we close the chain by connecting the d'th deme back to the first deme. As a consequence, all demes are fully equivalent. Biologically, this scenario corresponds, for example, to a chain of shallow-water habitats around an island or close to the shore of a lake. We also note that for very long chains we should expect that the influence of the boundary conditions becomes negligible.

The coalescence time for a pair of alleles will only depend on their initial distance. Accordingly, we denote with T_i the coalescence time of a pair with initial distance of i and $0 \leq i \leq d/2$. We find the following recursion

$$\mathrm{E}[T_0] = \frac{1 + 2M\mathrm{E}[T_1]}{d + 2M} \tag{1.26}$$

$$\mathrm{E}[T_i] = \frac{1}{2M} + \frac{1}{2}\Big(\mathrm{E}[T_{i+1}] + \mathrm{E}[T_{i-1}]\Big)\,,\ 0 < i \leq (d/2) - 1\,. \tag{1.27}$$

From the proposition above, we already know that $\mathrm{E}[T_0] = 1$ must hold. We thus have

$$\mathrm{E}[T_1] = 1 + \frac{d-1}{2M}$$

and in general

$$\mathrm{E}[T_k] = 2\mathrm{E}[T_{k-1}] - \mathrm{E}[T_{k-2}] - \frac{1}{M} = 1 + \frac{k(d-k)}{2M}\,, \tag{1.28}$$

which is easily proved by induction. Note that model and results reduce to the island model for $d \leq 3$. For a very long chain with $d \gg k$, we obtain $\mathrm{E}[T_k] \approx 1 + kd/(2M)$. The expected coalescence time thus increases approximately linearly with the distance. We can define a distance-dependent f_{ST} to compare demes that are separated by a fixed difference,

$$f_{ST}(k) = \frac{\mathrm{E}[T_k] - \mathrm{E}[T_0]}{\mathrm{E}[T_k] + \mathrm{E}[T_0]} = \frac{k(d-k)}{4M + k(d-k)}\,. \tag{1.29}$$

We can also derive an expected coalescent time for the total population,

$$\mathrm{E}[T_T] = 1 + \frac{\sum_{k=1}^{(d-1)/2} k(d-k)}{dM} = 1 + \frac{(d-1)(d+1)}{12M} \tag{1.30}$$

(for odd d) and thus

$$f_{ST} = \frac{\mathrm{E}[T_T] - \mathrm{E}[T_0]}{\mathrm{E}[T_T]} = \frac{(d-1)(d+1)}{12M + (d-1)(d+1)}\,. \tag{1.31}$$

Population structure across the whole habitat becomes irrelevant if $M \gg d^2$, which is in marked contrast to the symmetric island model for large d.

2-dim stepping-stone model

Results for a 2-dimensional stepping-stone model on a square lattice can also be found (see Durrett 2008), but all require lengthly calculations. There are two main approaches, one using Fourier methods (making use of the spatial homogeneity), the other one using probabilistic arguments. For the latter, consider the problem on the torus (square lattice with periodic boundary conditions) of size $d = L^2$. For a random pair of individuals, we can decompose the coalescence time into two parts,

$$T_T = T_s + T_0\,,$$

where the first part T_s measures the time to bring both individuals to the same deme, and a second part T_0 represents the time to coalescence once they are in the same deme (for the first time). As in the 1-dim case, we already know that $E[T_0] = 1$. For $E[T_s]$, we note that the migration rate M just scales the speed of the deme hopping, thus

$$E[T_s] = \frac{E[K_s]}{M}\,,$$

where K_s is the number of steps needed by a random walker to reach some deme 0 from a random deme on the $L \times L$ torus. Cox and Durrett (2002) show that $E[K_s] \sim L^2 \log[L] \sim d \log[d]$ for large $d = L^2$ (cf Durrett 2008, section 5.3.1). This shows that there are two different limits of interest:

1. For $M \gg d \log[d]$, the time for migration to the same deme is irrelevant and the positions of the demes for both individuals does not matter. We then have very weak structure, indicated by $f_{ST} = 0$, and the coalescent is equivalent to the one of a panmictic population.

2. For $M \ll d \log[d]$, we have strong population structure with $f_{ST} > 0$ and the time to coalescent is dominated by the random walk of genealogical lines on the torus.

Comparing the critical migration rates M_c that distinguish weak and strong structure for the 1-dim and 2-dim stepping stone models and the island model, we get $M_c \sim d^2$ (1 dim), $M_c \sim d \log[d]$ (2 dim), and $M_c \sim 1$ for the island model. It should be noted, however, that M_c measures migration between pairs of demes. The total migration rate per deme (to all other demes) $M_{c,\mathrm{tot}}$ differs from M_c only by a constant factor for the stepping stone models, but by a factor d for the island model. We thus have $M_{c,\mathrm{tot}} \sim d^2$ (1 dim), $M_{c,\mathrm{tot}} \sim d \log[d]$ (2 dim), and $M_{c,\mathrm{tot}} \sim d$ (island), demonstrating a strong effect of isolation by distance in one dimension, but a much weaker effect in two dimensions in the limit of large d.

Strong migration limit

For general (ergodic) migration schemes, the linear equation system for the coalescence time cannot usually be solved. We can make further progress, however, in the limit of strong migration (the first limiting case for the 2-dimensional stepping-stone model above). In this limit, we can assume that the system reaches the migration equilibrium (i.e., the stationary distribution of the backward migration process) before the first coalescent event. The coalescence probability of a pair then becomes independent of the initial state and can be expressed as

$$p_{c,1} = \sum_{i=1}^{d} \frac{u_i^2}{2N_i} = \frac{1}{2N_e^{(c)}}\,, \tag{1.32}$$

where u_i is the probability of deme i in the stationary distribution. $N_e^{(c)}$ is the coalescence effective size. The corresponding coalescence rate on the time scale of $2N$ generations is $\sum_i (u_i^2/c_i)$. For the coalescence time, we can ignore the time period until the migration equilibrium is reached. Thus

$$\mathrm{E}[T_S] = \mathrm{E}[T_T] = \left(\sum_{i=1}^{d} \frac{u_i^2}{c_i}\right)^{-1} = \left(\sum_{i=1}^{d} u_i \left(\frac{c_i}{u_i}\right)^{-1}\right)^{-1} < \sum_{i=1}^{d} u_i \frac{c_i}{u_i} = 1\,, \tag{1.33}$$

since the weighted harmonic mean is always smaller than the arithmetic mean. We thus see that population structure will usually lead to shorter coalescence times (or smaller effective population sizes) in the strong migration limit, unless migration is conservative, $u_i = c_i$ (like, for example, in the stepping-stone model). Of course, we always have $f_{ST} \to 0$ for strong migration.

1.5 General structured coalescent

It is straightforward to generalize the coalescent process in a structured population from just two alleles to a sample of size n. Assume that we have n_i sequences sampled from deme i and $\sum_i n_i = n$. There are two types of events in the genealogy of the sample:

1. On the coalescent time scale (in units of $2N$ generations), the rate of coalescence events in the ith deme is
$$\binom{n_i}{2}\frac{1}{c_i}\,, \tag{1.34}$$
2. On the same scale, the rate for backward migration from deme i to deme j is
$$2Nn_i m_{ij} =: \frac{1}{2} n_i M_{ij}\,, \tag{1.35}$$

where $M_i = 4Nm_i$ as before. We can now construct the coalescent process as a continuous-time Markov process on the state space of configurations with elements

$$\mathbf{n} = (n_1, n_2, \ldots, n_d), \qquad n_i \in \mathbb{N}_0 \tag{1.36}$$

with the transition rate matrix $\exp[\tau\mathbf{Q}]$ with

$$Q_{\mathbf{n},\mathbf{n}'} = \begin{cases} \binom{n_i}{2}\frac{1}{c_i} & ; \mathbf{n}' = \mathbf{n} - \mathbf{e}_i \\ n_i M_{ij}/2 & ; \mathbf{n}' = \mathbf{n} - \mathbf{e}_i + \mathbf{e}_j \\ -\sum_i \left(n_i M_i/2 + \binom{n_i}{2}\frac{1}{c_i} \right) & ; \mathbf{n} = \mathbf{n}' \\ 0 & ; \text{else} . \end{cases} \tag{1.37}$$

Here, $\mathbf{e}_i = (0, 0, \dots, 1, 0, \dots, 0)$ with entry 1 in the ith position denotes the ith unit vector. We can construct the genealogical process as follows:

1. For the process in state $\mathbf{n}$, the time $T_\mathbf{n}$ to the next event is exponentially distributed with parameter $(-Q_{\mathbf{n},\mathbf{n}})$, respectively, with expected value
$$\mathrm{E}[T_\mathbf{n}] = \frac{-1}{Q_{\mathbf{n},\mathbf{n}}} . \tag{1.38}$$
2. The probability that this event is a coalescence event in deme i, or a backward migration event from deme i to deme j is
$$\Pr[\text{coal. in } i|\mathbf{n}] = \frac{-\binom{n_i}{2}/c_i}{Q_{\mathbf{n},\mathbf{n}}} \quad \text{or} \quad \Pr[\text{mig. } i \to j|\mathbf{n}] = \frac{-n_i M_{ij}/2}{Q_{\mathbf{n},\mathbf{n}}} , \tag{1.39}$$
respectively. The distribution of topologies follows from a random choice of lines from deme i for either coalescence or migration.

Infinite islands model

It is even more difficult to obtain fully analytical results for a structured coalescent process of a sample of size n than for just a pair of individuals. Even for a symmetric island model, we need to distinguish a large number of states to account for the various ways how the lines can be distributed over the islands (for 4 lines there are already 5 states). There are two limits where the problem becomes manageable: one is the strong migration limit, where the problem reduces to the panmictic case (and thus the Kingman coalescent) with a changed effective population size $N_e^{(c)}$. We can derive $N_e^{(c)}$ whenever we can determine the stationary distribution of the backward migration process. The other limit is the co-called infinite islands model due to John Wakeley (1998). The general idea is as follows: consider the coalescent process with a sample taken from a finite number of demes. If the number of demes is very large, we can separate the the total time to the most recent common ancestor into two phases.

- During the first phase, called the *scattering phase* by Wakeley, lines in each local deme either coalesce or leave the deme by backward migration. We assume that every migration event leads to a previously unoccupied island. Consequently, we can also ignore immigration of lines into the focal local deme during this time.

- The second phase, called the *collection phase*, starts once there is no deme anymore with more than a single line. It describes the coalescent process from this starting condition, where events occur on a much larger time-scale as compared to the first phase. It is convenient to start the detailed discussion with the second phase.

The collecting phase We start our process with n individuals in different demes. For a very large number of demes, $d \gg n$, all ancestral lines will be in different demes for most of the time. Only occasionally, with probability $\sim (n/d)$, two lines will meet in the same deme. The probability that more than two lines meet in the same deme, or that there are two or more demes with more than a single line is $\sim (n/d)^2$ and can be ignored for $d \to \infty$. As a consequence, we only need to consider two states in the process: one with all lines in different demes (state 1) and a second one with two lines in one deme and all other lines in different demes (state 2). Call $\mathrm{E}[T_1]$ and $\mathrm{E}[T_2]$ the expected time to the next coalescent event in state 1 or 2, respectively. Then

$$\mathrm{E}[T_2] = \frac{1}{d+(d-1)M} + \frac{(d-1)M}{d+(d-1)M}\,\mathrm{E}[T_1] \quad \xrightarrow{d\to\infty} \quad \frac{M}{1+M}\,\mathrm{E}[T_1]\,, \tag{1.40}$$

$$\mathrm{E}[T_1] = \frac{1}{(n-1)nM/2} + \mathrm{E}[T_2] = \frac{1}{(n-1)nM/2} + \frac{M}{1+M}\,\mathrm{E}[T_1]\,. \tag{1.41}$$

We thus obtain

$$\mathrm{E}[T_1] = \frac{1+M}{(n-1)nM/2} = \left(1+\frac{1}{M}\right)\binom{n}{2}^{-1} \tag{1.42}$$

on a scale of $2N$ generations. Since each pair of lines will coalesce with the same probability, this is just the normal Kingman coalescent on a larger time scale.

The scattering phase The coalescent process during this phase corresponds to the so-called *coalescent with killings*, where lines can either coalesce or vanish (get “killed”) by time-backward emigration. The problem is entirely analogous to the statistics of the number of haplotypes for a standard neutral coalescent process and mutation according to the infinite alleles model (where lines are “killed” by the mutation events, see lecture *Mathematical Population Genetics*). The results therefore directly carry over, with the population mutation parameter $\theta = 4Nu$ replaced by the migration parameter $M = 4Nm$. In particular, the number K_n of alleles in a sample of size n from a single deme that go back to different migration events is given by the Ewens' sampling distribution,

$$P[K_n = k] = \frac{M^{k-1}}{(M+1)(M+2)\cdots(M+n-1)} \cdot S(n,k)\,, \tag{1.43}$$

where the $S(n,k)$ denotes the Stirling numbers of the second kind. The distribution of size classes of alleles that go back to each migration event is given by the other part of Ewens' formula,

$$P_n[a_1,\ldots,a_k|K_n = k] = \frac{n!}{S(n,k)}\prod_{j=1}^{n}\frac{1}{a_j!\,j^{a_j}}\,, \tag{1.44}$$

where a_j, $1 \leq j \leq k$ denotes the size of the jth class and $\sum_j a_j = n$. The total time for the scattering phase is only $\sim N/d$ and can be ignored relative to the time spent during the collecting phase.

1.6 Structure and pattern

So far, we have discussed the effect of population structure on an expected neutral genealogy of a population sample. This is, we have been concerned with *descent*. We have used two main measures. First, we have measured the effect of population structure on the effective population size $N_e^{(c)}$ that defines the time scale on which coalescence happens. Here, population structure is only one of many factors that can lead to an altered $N_e^{(c)}$ – and some patterns of (even strong) population structure do not affect the effective population size. Second, we have defined the probabilities for identity by descent within a group f_i and the ratios of the coalescence times within and among groups, f_{IT} or f_{ST}. Any value of f_{IT} or $f_{ST} > 0$ means that individuals (genes) in the total population are not all exchangeable.

All our measures of population structure so far did not use the allelic *state* for their definition. This is appropriate: after all, the presence or absence of a polymorphism depends on many factors (such as the demographic history or the mutation rate) that should not decide whether a population is more or less structured. However, the reverse direction is certainly true: population structure can (and usually will) influence the typical polymorphism pattern. We can thus hope that we can at least obtain some information about the structure from the pattern. This is of great practical relevance since polymorphism patterns (other than genealogies) are directly observable. To understand the consequences of population structure for polymorphism patterns, we need to add mutation to the genealogical process. As we will see below, this can be done in several ways and the result depends on the mutation model that is used.

Heterozygosities and fixation indices

Before we discuss specific results, we will first describe the measures that are used to characterize the polymorphism pattern in a structured population. By far the most widely used measure is the heterozygosity H or its complement, the homozygosity $F = 1 - H$. The homozygosity measures *identity by state*. As such, it is an observable that is closely related to *identify by descent* - which is usually not an observable quantity. Similarly, heterozygosity measures difference in state. We have previously (in the lecture *Mathematical Population Genetics*) defined H as the probability that two homologous alleles from the population are different. In a structured diploid population this definition can be refined:

- On the highest level, the *total heterozygosity* H_T is defined as the probability that two randomly chosen alleles from the entire population are different (by state).
- On an intermediate level, we compare two randomly drawn alleles from the same (randomly chosen) subpopulation or deme. They are different with probability H_S,

the *subpopulation heterozygosity.*

- On the lowest level, we define the *individual heterozygosity* H_I as the probability that the two homologous alleles of a single, randomly chosen individual are different.

With three different levels of heterozygosity, we can define three fixation indices or so-called F-statistics (first introduced by Sewall Wright),

$$F_{IS} = \frac{H_S - H_I}{H_S}, \quad F_{IT} = \frac{H_T - H_I}{H_T}, \quad F_{ST} = \frac{H_T - H_S}{H_T}. \tag{1.45}$$

F_{IS}, F_{IT}, and F_{ST} detect differences in genetic variation due to non-random mating at different levels. They are related as

$$(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST}) \tag{1.46}$$

In words: *The proportion of variation in the total population due to variation within an individual equals the proportion of variation in a subpopulation due to variation within an individual times the proportion of variation in the total population due to variation in the subpopulation.*

- The relation reflects the hierarchical structure of allelic groups: the two alleles in an individual (a minimal family group) are always in the same neighborhood (or spatial group).

- (1.46) implies that any two of the three measures contain the full information. The commonly used measure to describe spatial population structure is F_{ST}, i.e., the proportion of genetic variation among individuals drawn from all subpopulations that is due to genetic differences between subpopulations. The commonly used measure to describe inbreeding is F_{IS} (if there is spatial structure $F_{ST} > 0$, otherwise $F_{IT} = F_{IS}$ are equivalent).

We can summarize several further basic properties of the fixation indices that hold independently of a mutation model. Assume that the total population is partitioned into d disjoint demes, where $c_i = N_i/N$ is the relative size of the ith deme, $\sum_i c_i = 1$. Consider a single locus with two alleles A and a. Let p_i be the frequency of A in deme i.

- $\bar{p} = \sum_i c_i p_i$ is the frequency of A in the total population and we derive

$$H_T = 2\bar{p}(1 - \bar{p}) = 2\sum_i c_i p_i \Big(1 - \sum_i c_i p_i\Big) \quad ; \quad H_S = \sum_i c_i 2 p_i (1 - p_i)\,. \tag{1.47}$$

Thus

$$H_T - H_S = 2\sum_i c_i p_i^2 - 2\Big(\sum_i c_i p_i\Big)^2 = 2\sum_i c_i \Big(p_i - \bar{p}\Big)^2 = 2\mathrm{Var}[p_i] \tag{1.48}$$

is just the variance of the allele frequency across demes (weighted by the deme size) and

$$F_{ST} = \frac{\mathrm{Var}[p_i]}{\bar{p}(1-\bar{p})} > 0\,. \tag{1.49}$$

This relation is sometimes used as an alternative definition of F_{ST}. It works for a biallelic locus .

- Equation (1.49) shows that 0 is a lower bound for F_{ST}, which corresponds to no differences in the heterozygosity among subpopulations and thus reflects absence of population structure. Note that this means that H_S cannot be larger than H_T. The opposite extreme, $F_{ST} = 1$, is reached if the total population is composed out of homogeneous subpopulations (i.e., $H_S = 0$). It reflects complete population structure.

- Also for F_{IT} and F_{IS} the maximal value is 1, which is reached when all individuals are homozygous ($H_I = 0$). Hardy-Weinberg equilibrium in a subpopulation is indicated by $H_S = H_I$ and thus $F_{IS} = 0$. In the opposite extreme, if all individuals are heterozygous, we have $H_I = 1$ and $p_i = 0.5$ for a biallelic locus. Consequently, $H_S = H_T = 0.5$ and $F_{IS} = F_{IT} = -1$. In contrast to F_{ST}, we thus find that F_{IS} and F_{IT} can become negative. Technically, this difference is caused by a slight difference in the sampling scheme that is used in both cases: Sampling is done with replacement for H_S and H_T, i.e., the probability for an A allele in the second draw does not depend on the result of the first draw. In contrast, sampling for H_I is always without replacement. In this case, if all individuals are heterozygous, the draw of the first allele will determine the state of the second allele.

- Eq. (1.49) shows that F_{ST} is larger than zero as soon as there are any deviations of the allele frequencies among demes. Assume that all subpopulations are in Hardy-Weinberg equilibrium and thus $F_{IS} = 0$. From (1.46) we see that nevertheless $F_{IT} = F_{ST} > 0$ in this case. I.e. there is an excess of homozygous individuals if we take a sample from the total population. This is called the *Wahlund effect* (Wahlund 1928): population structure has similar effects on F_{IT} as preferred mating among relatives. For this reason, populations need to be assessed for hidden spatial structure before an observation of $F_{IT} > 0$ can be interpreted as evidence for preferred mating among relatives.

For an inbreeding model, we can relate F_{IT} to the inbreeding coefficient f. To do this, assume that we have two alleles segregating in the population at a given locus, A and a, with frequencies p and $q = 1 - p$, respectively. We obviously have $H_T = 2pq$. For a population in Hardy-Weinberg equilibrium, also $H_I = 2pq$, but for an inbreeding population, H_I will deviate from this value. Imagine a population with inbreeding coefficient f. We can argue as follows. If the total population size is much larger than the size of a family group, coalescence within a family will typically happen quickly (if it happens). We can then ignore the rare event of a mutation on this part of the genealogy. In this case,

$$H_I = (1 - f)H_T \tag{1.50}$$

and

$$F_{IT} = \frac{H_T - H_I}{H_T} = f\,. \tag{1.51}$$

The same holds for populations with additional spatial substructure, where H_T is replaced by H_S. We obtain the equilibrium genotype distribution for inbreeding in a (sub)population as

$$\begin{array}{rll} p^2 + fpq & \text{for genotype} & AA, \\ 2(1-f)pq & \text{for genotype} & Aa, \\ q^2 + fpq & \text{for genotype} & aa. \end{array}$$

We see that F_{IT} just coincides with f in this case: Identity by descent (ibd) implies identity by state in this approximation.

- Since F_{IT} can readily be measured form data, it is frequently used as a proxy for f, the level of inbreeding. Sometimes, (1.51) is even used as the definition of f. However, this is generally not appropriate. Indeed, there can be other causes for the distortion of Hardy-Weinberg proportions (such as selection) that have nothing to do with inbreeding.

- As mentioned above, selection or disassortative mating can also lead to $H_I > H_T$ and a negative F_{IT}. This, once again, highlights the difference to the inbreeding coefficient f, which has been defined as a probability and is thus bound to be positive.

- With inbreeding, deviations from Hardy-Weinberg proportions can build up over many generations and the equilibrium $F_{IT} = f$ is usually only reached in the limit $t \to \infty$ (with the exception of the most basic sperm-egg association model). In contrast, it only takes a single generation of random mating to return to Hardy-Weinberg proportions.

- Finally, note that inbreeding as such does not change the allele frequencies (it does not induce selection), but will only regroup these alleles in the genotypes of individuals and of family groups.

Analogous results hold for spatial structure and F_{ST} if the expected time to the first event within a subpopulation is sufficiently short that mutations can be ignored during this time. This typically holds true for the infinite-islands model.

Mutation models: Infinite alleles and infinite sites

In our treatment so far, we have expressed the fixation indices as a function of the allele frequencies across demes. This leaves the question open what these allele frequencies are in the first place, for a given model of population structure. To capture population structure with fixation indices, we therefore need to derive these measures directly from the model parameters. To this end, we need to add mutation to our framework. In discrete generations, one usually assumes that mutations occur with a certain probability in every newborn individual. In the coalescent setting, we can simply assume that mutations occur with a constant rate $\Theta/2 = 2Nu$ along each line of descent.

Infinite alleles. In the infinite alleles model, each mutation leads to a novel allele. Two individuals will therefore be identical in state if and only if there is no mutation in their genealogy up to their most recent common ancestor. If T is the time to the most recent common ancestor, the number of mutations on the genealogy is Poisson distributed with parameter ΘT. The Poisson probability for at least a single mutation, given T, is $1 - \exp[-\Theta T]$. Averaging over T, we thus obtain in the panmictic case

$$H = \mathrm{E}\Big[1 - \exp[-\Theta T]\Big] = \int_0^\infty \Big[1 - \exp[-\Theta T]\Big] \exp[-T]\, dT = \frac{\Theta}{\Theta + 1}\,,$$

since T is exponentially distributed with parameter 1 (in coalescent scaling). For the structured coalescent, we can make use of the fact that times to consecutive events in the genealogy of the sample are independently exponentially distributed. If λ_0 is the total rate of events that can happen in the initial phase, we can express H as

$$H = 1 - \mathrm{E}\Big[\exp(-\Theta T)\Big] = 1 - \int_0^\infty \exp[-\Theta T_0]\lambda_0 \exp[-\lambda_0 T_0]\, dT_0 (1 - \tilde{H}) = \frac{\Theta + \lambda_0 \tilde{H}}{\Theta + \lambda_0}\,,$$

where $\tilde{H}$ denotes the expected heterozygosity after the first event. For a structured population with discrete demes, as introduced above, we denote the expected heterozygosity of two alleles in deme i and deme j as H_{ij}. We then find the following recursion:

$$H_{ii} = \frac{\Theta + \sum_{j \neq i} M_{ij} H_{ij}}{\Theta + (1/c_i) + \sum_{j \neq i} M_{ij}} \tag{1.52}$$

$$H_{ij} = \frac{2\Theta + \sum_{k \neq i} M_{ik} H_{kj} + \sum_{k \neq j} M_{jk} H_{ik}}{2\Theta + \sum_{k \neq i} M_{ik} + \sum_{k \neq j} M_{jk}}\,. \tag{1.53}$$

For the island model, in particular, this simplifies to

$$H_S = \frac{\Theta + (d-1) M H_B}{\Theta + d + (d-1)M} \tag{1.54}$$

$$H_B = \frac{\Theta + (d-2) M H_B + M H_S}{\Theta + (d-1)M}\,. \tag{1.55}$$

The second equation can be rewritten as

$$H_B = \frac{\Theta + MH_S}{\Theta + M}$$

and we obtain the solution

$$H_S = \Theta \, \frac{\Theta + dM}{(\Theta + d + Md)\Theta + dM} \,, \tag{1.56}$$

$$H_B = \Theta \, \frac{\Theta + d + dM}{(\Theta + Md + d)\Theta + dM} = H_S + \frac{\Theta}{M + (1+M)\Theta + \Theta^2/d} \tag{1.57}$$

and

$$\begin{aligned} H_T &= \Theta \, \frac{\Theta + dM}{(\Theta + d + Md)\Theta + dM} + \frac{d-1}{d} \frac{\Theta}{M + (1+M)\Theta + \Theta^2/d} \\ &= \Theta \, \frac{\Theta + dM + d - 1}{(\Theta + d + Md)\Theta + dM} \,. \end{aligned} \tag{1.58}$$

This results in an F_{ST} of

$$F_{ST} = \frac{d-1}{\Theta + dM + d - 1} = \frac{1}{1 + Md/(d-1) + \Theta/(d-1)} \,. \tag{1.59}$$

Infinite sites. For mutations at a single nucleotide position, we usually have $\Theta \ll 1$. This is the basis of the *infinite sites model* (ISM), where we assume that a single nucleotide position is hit by a mutation at most once. In this limit, we find that the heterozygosities are linear functions in the mutation parameter, i.e.,

$$H_{ij}^{(\mathrm{ISM})} = \Theta \, \frac{\partial}{\partial \Theta} H_{ij}(\Theta) \Big|_{\Theta=0} = \Theta \, \frac{\partial}{\partial \Theta} \mathrm{E}\Big[1 - \exp[-\Theta T_{ij}]\Big]_{\Theta=0} = \Theta \, \mathrm{E}[T_{ij}] \,, \tag{1.60}$$

where $\mathrm{E}[T_{ij}]$ is the expected coalescence time of an ij pair. Thus also

$$F_{ST}^{(\mathrm{ISM})} = \frac{H_T^{(\mathrm{ISM})} - H_S^{(\mathrm{ISM})}}{H_T^{(\mathrm{ISM})}} = \frac{\mathrm{E}[T_T] - \mathrm{E}[T_S]}{\mathrm{E}[T_T]} = f_{ST} \,. \tag{1.61}$$

For the infinite sites model (and more generally in the limit of low mutation), the state-based fixation index F_{ST} thus coincides with the descent-based measure f_{ST} for population subdivision that we have defined earlier.

In general, we observe the following:

- F_{ST} is a monotonically decreasing function of the mutation parameter Θ. For small Θ, F_{ST} captures the population structure as measured by f_{ST}. For larger Θ, mutations diffuse this signal. For the island model, we see that this happens once Θ is of the order of dM. In general, we can reason as follows: If $\Theta > \sum_j M_{ij}$ for some deme i, then a line of descent started in this deme will likely mutate before it migrates to another deme. However, in the infinite alleles model, the history prior to the first mutation event is irrelevant for the observed pattern (we can stop the line like in the coalescent with killings). We thus see that the pattern becomes insensitive of the migration pattern of the model. Consequently, the level of migration cannot be estimated from polymorphism patterns.

- A consequence of the observation above is that F_{ST} is not a good measure of spatial population structure if the differentiation is very strong, and thus $M \ll 1$. To illustrate this point, consider a structured population with a large number of demes. An arbitrary fraction of these demes is fixed for allele a and all others are fixed for allele A (at least one deme for each allele). Then $H_S = 0$ and $H_T > 0$, thus $F_{ST} = 1$, although there is no differentiation at all among most demes. We get the same result as if each deme was fixed for a different allele.

- Alternatively, consider a scenario with very many alleles at a given locus. Assume that we have complete differentiation across d demes of equal size, such that different demes do not share any alleles at the locus. Then $H_T = (d - 1 + H_S)/d$ and $F_{ST} = (d-1)(1-H_S)/(d-1+H_S)$. Thus $F_{ST} = 1$ if $H_S = 0$, but $F_{ST} < 1$ for $H_S > 0$, and even $F_{ST} \to 0$ for $H_S \to 1$, although there is complete differentiation among demes.

- We conclude that F_{ST} is a meaningful measure of differentiation only if the level of differentiation and diversity is low (i.e., for small Θ and large M). For high diversity levels, alternative measures for population differentiation have been suggested. For example, we can consider the ratio of the homozygosities instead of the heterozygosities. For the island model,

$$D = \frac{1 - H_S}{1 - H_T} = \frac{d\Theta + dM}{\Theta + dM}\,. \tag{1.62}$$

  For large M/Θ, D converges to 1 and is not informative. For small M/Θ, it approaches the number of demes d. In general, we find $D \to 1/\sum_i c_i^2$ in this limit, which is the inverse of the probability that two random individuals are from the same deme. Note that this quantity does indeed provide some information about population structure. However, it is independent of the migration rate (which therefore cannot be estimated). Note that D does provide some information in both special scenarios described above, unless $H_S \to 1$.

We note that explicit expressions for the expected heterozygosity (or homozygosity) can be used to derive moments of the distribution of coalescence times (formally, $1-H(-\Theta)$ is the moment generating function or *Laplace transform* of the coalescence time). For example, in the case of the island model,

$$\mathrm{E}[T_S] = \frac{\partial}{\partial\Theta} H_S\Big|_{\Theta=0} = 1\,, \tag{1.63}$$

$$\mathrm{Var}[T_S] = \frac{-\partial^2}{\partial\Theta^2} H_S\Big|_{\Theta=0} - 1^2 = 2\,\frac{dM + d - 1}{dM} - 1 = 1 + \frac{2(d-1)}{dM}\,, \tag{1.64}$$

confirming our earlier results.

Estimating F_{ST}

We have defined the heterozygosities H_S and H_T as expected values for a given model. From real data, we obtain estimators of these quantities. These estimators should average over all sources of stochasticity, in particular due to mutation and the coalescent history. This is best done by taking samples from many loci across the whole genome. On the level of single nucleotide polymorphisms (SNPs), heterozygosity is also called *nucleotide diversity* and denoted as π. To estimate π, we simply average the pairwise sequence differences across the total sequenced region. For a panmictic population, the corresponding estimator $\hat{\pi}$ is a summary statistic of the site frequency spectrum. In a sample of size n,

$$\hat{\pi} = \frac{1}{L\binom{n}{2}} \sum_{i=1}^{n-1} i(n-i)\xi_i \,, \tag{1.65}$$

where L is the length of the sequence and ξ_i the number of polymorphisms with frequency (i/n) of the derived allele (mutation of size i).

In a structured population, we can express the heterozygosities H_T and H_S as summary statistics of the *joint site-frequency spectrum.* Suppose that we sample n_k alleles from deme k and $n = \sum_k n_k$. Also suppose that we have a balanced sample, such that $n_k/n = N_k/N = c_k$. Let $\xi_{i_1,\ldots,i_d}$ be the number of polymorphic sites, where the derived allele appears i_k times in deme k, $1 \le k \le d$ and $0 \le i_k \le n_k$. Then

$$\hat{\pi}_T = \frac{1}{L\binom{n}{2}} \sum_{i_1,\ldots,i_d} \xi_{i_1,\ldots,i_d} \Big(\sum_k i_k\Big)\Big(n - \sum_k i_k\Big) \tag{1.66}$$

and

$$\hat{\pi}_S = \sum_k c_k \hat{\pi}_{kk} = \sum_k \frac{c_k}{L\binom{n_k}{2}} \xi_{i_1,\ldots,i_d}\, i_k(n_k - i_k) \,. \tag{1.67}$$

Given polymorphism data from a structured population, we can proceed as follows:

- Measure (estimators for) global an local heterozygosities.
- Test (globally and pairwise) whether H_S and H_T are significantly different.
- If yes: derive (estimators for) F_{ST}.
- See whether there is evidence for isolation by distance. If not: possibly use the island model to estimate migration rates. Otherwise: try to fit more complex models (this may require summary statistics other than F_{ST}).

2 Selection footprints

We have seen how genealogies of population samples are affected by population demography (lecture *Mathematical Population Genetics*) and population structure (in the first part of this lecture). So far, however, we have exclusively been concerned with neutral evolution. We'll now ask for the impact of natural selection. Doing so, we immediately face a problem: coalescent theory does not easily generalize to selection models, where the state of alleles with different fitness will affect the genealogical relationships. In particular, it is no longer possible to construct the descent of an allele without the knowledge about its state. There are two main ways how to proceed.

1. The ancestral selection graph (Neuhauser and Krone) proceeds by first producing a graph that contains all *potential* genealogies of a biallelic locus under selection. In a second step, the allelic state is decided and the actual genealogy is deduced.

2. An alternative approach is to determine the allele frequencies of the beneficial allele through time in a first step (e.g. using a forward-in-time formalism). In a second step, we can then model the genealogy of a population sample backward in time conditioned on the trajectory of the beneficial allele. This second step makes use of the structured coalescent.

2.1 The ancestral selection graph

We will briefly describe the main idea of the ancestral selection graph. Assume (for simplicity) a haploid population of size $2N$ and a single locus under selection with two alleles A and a. There is symmetric mutation at rate u from A to a and vice-versa. In the absence of selection, we obtain a standard neutral (Kingman) coalescent, where the genealogy can be constructed first and mutations (at rate $\Theta/2 = 2Nu$) can be added in a second step. Let us now assume that A and a differ in fitness, with an advantage of A that is given by a selection coefficient $s > 0$. We can easily include selection forward in time. In a continuous time framework (e.g., a Moran model), this can be done by assuming that A individuals are subject to "extra birth" events at a constant rate s, where a single A offspring is produced and replaces a randomly chosen individual from the population. In a Wright-Fisher model, we can similarly assume that after each round of neutral reproduction all A individuals additionally reproduce with probability s and replace randomly chosen types.

Let us now see how this process looks like in the backward (coalescent) direction. In particular, we want to record the extra birth events in the genealogy. In the coalescent scaling, we let $s \to 0$ and $N \to \infty$ such that $\alpha = 2Ns$ assumes a finite value. With this scaling, we can, once again, ignore double events at a single time point. In particular, the probability that coalescence occurs directly at an extra birth event is $\sim s/N = \alpha/(2N^2)$ per generation, or $\sim \alpha/N \xrightarrow{N\to\infty} 0$ in coalescent time units. Assume first that all individuals are of A type (no mutation at the selected locus). We then simply have extra birth events that occur with rate α for each lineage, independently of the coalescence events. Of course, nothing of genealogical relevance happens at these events: alleles do not coalesce and they do not change their type. Marking extra births is thus just an accounting exercise.

Now assume that both alleles A and a segregate in the population. Forward in time, only A alleles have extra birth events. Backward in time, the rate of extra birth depends on the frequency p_A of the allele in the population, which is not known. The crucial idea of the ancestral selection graph is that we still assume a constant extra birth rate of α, as if any individual in the population could give extra birth. However, we also account for the possibility that this extra birth is not real. We proceed in the following steps:

1. Potential extra birth events happen at rate α per lineage. At each such event, we assign two potential ancestors: one is called the *incoming line*. This will be the real ancestor if and only if it has type A (and thus the capacity for extra birth). The other one is called the *continuing line* and will be the ancestor if the incoming line is not of type A. Since we do not know the state of the incoming line, we follow both possibilities.

2. Backward in time, we thus obtain *splitting events* at rate α for each line. If there are currently j lines, we have a total rate of αj for splitting and a rate of $j(j-1)/2$ for coalescence. Instead of a simple tree, we obtain a graph that is generated by a special type of a birth-death process. The important point is that we can generate this graph without knowledge of the allelic state or the frequency of the A allele in the population.

3. We follow this graph to the first time where only a single line exists. This is called the *ultimate ancestor* (UA). Since the splitting rate is linear in the number of lines, but the branching rate is quadratic, an ultimate ancestor will be reached in finite time with probability one. However, as we will see below, this time can be very long unless selection is very weak.

4. We can also place mutations on the graph with a rate of $\Theta/2$ along each branch. At a mutation, the state of the allele will change its type, either from A to a or from a to A. (This simple procedure makes use of the symmetry of the mutation rates, but also asymmetric mutation can be included: we assign potential mutations at a fixed rate u that lead to states A and a with probabilities q_a and $(1-q_a)$. Only mutations that change the type will be effective mutations, which we can see after assigning the state to all ancestral lines).

5. We now need to decide on the state of the ultimate ancestor. This can be done by drawing from an equilibrium distribution. If we assume mutation-selection-drift balance, this equilibrium can be derived from diffusion theory: the frequency x of allele A has the density (cf. lecture *Mathematical Population Genetics*)

$$f(x) = Cx^{\Theta-1}(1-x)^{(\Theta-1)}\exp[\alpha x]\,. \tag{2.1}$$

6. In the last step, the states of all branches in the graph are decided and the real ancestors at the branching (extra birth) events are resolved. As a result, the real MRCA is usually reached much earlier that the UA. We now have generated a particular representative from the distribution of coalescent histories with selection. If we want, we can add neutral mutations to the genealogy (at sites in tight linkage to

the selected site). This is done by mutation dropping as in the case of the neutral coalescent.

To get a better intuition of the ancestral selection graph, we add a few observations:

- The ancestral selection graph makes use of the fact that we can extend a neutral genealogy to a larger sample size by just adding extra lines and letting them coalesce into the genealogy of the smaller sample. At each splitting event, an extra individual is added – but we only decide later, which one is the real ancestor and which one the added line.

- For strong selection, every line will carry multiple "extra birth" events and it is very likely that these are real ones. In this case, the state of an individual will mostly depend on whether there is a mutation event from A to a prior to the first split event (in backward direction). We then obtain the probability $\Theta/(\Theta+2\alpha) \approx \Theta/(2\alpha) = u/s$ for type a, consistent with the prediction of mutation-selection balance.

The so-called "size process" of the ancestral selection graph just counts the number of lines contained in the graph at a given time, before allelic states are added. Among other things, this size process determines the expected time to the ultimate ancestor. This is a relevant quantity, for example, if we want to estimate the computational costs for simulations of the ancestral selection graph. We prove the following theorem:

Theorem: Time to the ultimate ancestor *The expected time $T_{UA}(i)$ to the ultimate ancestor of an ancestral selection graph with split rate α per line and started with i lines is*

$$\mathrm{E}[T_{\mathrm{UA}}(i)] = 2\sum_{k=2}^{i}\sum_{\ell=0}^{\infty}\frac{(k-2)!(2\alpha)^{\ell}}{(\ell+k)!}\,. \tag{2.2}$$

Proof Mathematically, the size process is a continuous-time birth-death process with constant birth rate α per line and size-dependent death rate $(j-1)/2$ (per line if there are j lines). We first consider the embedded jump chain of this process, i.e., the discrete-time Markov chain that describes the state directly after each event, birth or death. The state space of this process is $j = 1, 2, 3, \ldots$ and transitions can only occur among neighboring states. An advantage of this simple structure of the transition matrix is that the process can be mapped to a martingale by a simple transformation and we then can make use of the strong properties of martingales. Since this is a frequent situation in population genetic models, we will demonstrate the method for a general case.

Consider a Markov chain with random variable X_t and state space $j = 1, 2, \ldots$. Transitions only occur among neighboring states with probabilities $P_{j\to j+1}$ and $P_{j\to j-1} = 1 - P_{j\to j+1}$. We can then define a corresponding martingale with a random variable $Y_t = \phi(X_t)$. The function $\phi(j)$ needs to fulfill the martingale property of a constant expectation, i.e.,

$$\mathrm{E}[Y_t|Y_0 = \phi(j)] = \mathrm{E}[Y_t|X_0 = j] = \phi(j)$$

for all j. This is the case if and only if

$$\phi(j) = P_{j\to j+1}\,\phi(j+1) + P_{j\to j-1}\,\phi(j-1)\,. \tag{2.3}$$

We are free to choose $\phi(1) = 0$ and $\phi(2) = 1$ and obtain the iteration

$$\phi(j+1) - \phi(j) = \frac{P_{j\to j-1}}{P_{j\to j+1}}\Big(\phi(j) - \phi(j-1)\Big) = \prod_{k=2}^{j} \frac{P_{k\to k-1}}{P_{k\to k+1}}\,,$$

and thus

$$\phi(j) = 1 + \sum_{\ell=2}^{j-1} \prod_{k=2}^{\ell} \frac{P_{k\to k-1}}{P_{k\to k+1}}\,.$$

Denote $t^*_{ij} = \min[t : X_t = j | X_0 = i]$ the first hitting time of state j for a process starting in state i. For $1 \le k < i < j$ and all transition probabilities $P_{\ell\to\ell\pm1} > 0$ for $\ell > 1$ we have $\min[t^*_{ij}, t^*_{ik}] < \infty$. For the martingale then follows

$$\begin{aligned}\phi(i) &= \mathrm{E}[Y_t | X_0 = i] \\ &= \Pr[t^*_{ij} < t^*_{ik}]\,\mathrm{E}[Y_t | X_{t^*_{ij}} = j, X_0 = i] + \Pr[t^*_{ij} > t^*_{ik}]\,\mathrm{E}[Y_t | X_{t^*_{ik}} = k, X_0 = i] \\ &= \Pr[t^*_{ij} < t^*_{ik}]\,\phi(j) + (1 - \Pr[t^*_{ij} < t^*_{ik}])\,\phi(k) \end{aligned} \tag{2.4}$$

and hence

$$\Pr[t^*_{ij} < t^*_{ik}] = \frac{\phi(i) - \phi(k)}{\phi(j) - \phi(k)}\,.$$

With $\phi(1) = 0$ we get for $j > i$

$$\Pr[t^*_{ij} < t^*_{i1}] = \frac{\phi(i)}{\phi(j)}\,.$$

Now assume that $\phi(j) \to \infty$ for $j \to \infty$ (or, more generally, that the $\phi(j)$ are unbounded). Then we find for $k \le i$

$$\Pr[t^*_{ik} < \infty] = 1\,.$$

In particular, also the state $k = 1$ will be reached in finite time with probability 1. Define the number of visits to state j before state $k = 1$ is reached from starting in state i as $n_j(i)$. Assume first that the process starts in j. The probability to never return to state j before state 1 is reached (thus, the total number of visits is $n_j(j) = 1$) is to go to $j - 1$ in the first step and then reach state 1 first, thus,

$$\Pr[n_j(j) = 1] = P_{j\to j-1}\Big(1 - \frac{\phi(j-1)}{\phi(j)}\Big)\,.$$

Assume now that the process starts in state i. If the process ever reaches state j then the number of visits to state j, will be geometrically distributed with mean $1/\Pr[n_j(j) = 1]$, thus

$$\mathrm{E}[n_j(i)] = \frac{\Pr[t^*_{ij} < t^*_{i1}]}{\Pr[n_j(j) = 1]} = \begin{cases} \frac{\phi(j)}{P_{j\to j-1}\big(\phi(j) - \phi(j-1)\big)} & ; \quad j \le i \\ \frac{\phi(i)}{P_{j\to j-1}\big(\phi(j) - \phi(j-1)\big)} & ; \quad j > i \end{cases}\,. \tag{2.5}$$

Returning to the ancestral selection graph, we have

$$j \to j+1 \quad \text{at rate} \quad \alpha j \quad ; \qquad P_{j\to j+1} = \frac{2\alpha}{2\alpha + j - 1}\,, \tag{2.6}$$

$$j \to j-1 \quad \text{at rate} \quad j(j-1)/2 \quad ; \qquad P_{j\to j-1} = \frac{j-1}{2\alpha + j - 1}\,. \tag{2.7}$$

For the martingale transformation of the jump chain, we have $\phi(1) = 0$ and $\phi(2) = 1$ and

$$\phi(i) = 1 + \sum_{\ell=2}^{j-1} \prod_{k=2}^{\ell} \frac{P_{k\to k-1}}{P_{k\to k+1}} = 1 + \sum_{j=2}^{i-1} \frac{(j-1)!}{(2\alpha)^{j-1}} = \sum_{j=2}^{i} \frac{(j-2)!}{(2\alpha)^{j-2}}\,. \tag{2.8}$$

We clearly have $\phi(i) \to \infty$ for $i \to \infty$. Using our result (2.5) for n_j above, we can now obtain the total time $T_j(i)$ that is spent with j ancestral lines before the ultimate ancestor (= state 1) is reached in the original continuous-time process started with i lines. Noting that the expected waiting time to leave a state is given by the inverse of the total rate, we have

$$\mathrm{E}[T_j(i)] = \frac{\mathrm{E}[n_j(i)]}{j(\alpha + (j-1)/2)} = \begin{cases} \frac{2\sum_{k=2}^{j}(k-2)!(2\alpha)^{j-k}}{j!}; & j \le i \\ \frac{2\sum_{k=2}^{i}(k-2)!(2\alpha)^{j-k}}{j!}; & j > i\,. \end{cases} \tag{2.9}$$

Finally, the total expected time to the ultimate ancestor derives as

$$\mathrm{E}[T_{\mathrm{UA}}(i)] = \sum_{j=2}^{\infty} \mathrm{E}[T_j(i)] = \sum_{j=2}^{\infty} \sum_{k=2}^{\min[i,j]} \frac{2(k-2)!(2\alpha)^{j-k}}{j!} \tag{2.10}$$

$$= 2\sum_{k=2}^{i} \sum_{j=2}^{\infty} \frac{(k-2)!(2\alpha)^{j-k}}{j!} \tag{2.11}$$

$$= 2\sum_{k=2}^{i} \sum_{\ell=0}^{\infty} \frac{(k-2)!(2\alpha)^{\ell}}{(\ell+k)!}\,, \tag{2.12}$$

which proves the theorem. For $\alpha = 0$, this reduces to the result of the Kingman coalescent

$$\sum_{k=2}^{i} \frac{2}{k(k-1)} < 2\,,$$

but for $\alpha > 0$, we have

$$\frac{\exp[2\alpha] - 1 - 2\alpha}{2\alpha^2} = \sum_{\ell=0}^{\infty} \frac{2(2\alpha)^{\ell}}{(\ell+2)!} \le \mathrm{E}[T_{\mathrm{UA}}(i)] \le 2(i-1)\exp[2\alpha] < \infty\,.$$

We thus see that the expected time to the ultimate ancestor is always finite, but increases exponentially with α.

2.2 The conditioned coalescent

The ancestral selection graph offers a rigorous framework to include selection into the coalescent process. However, it has several disadvantages. First, it is computationally inconvenient for strong selection, which is the most interesting case biologically. And second, it still requires the knowledge of the state of the ultimate ancestor (or, alternatively, of all lines in the graph at some other time). For an equilibrium distribution, we can (sometimes) determine this state by drawing from a known distribution. However, in many cases we are interested in non-equilibrium scenarios.

The key idea of the ancestral selection graph has been to construct an extended graph of equivalent lines in a first step, decide on the state later on, and deduce the real genealogy in a final step. With the conditioned coalescent, we take the opposite route. We first construct the frequency path of the selected allele in the population. This way, we distinguish two (or more) classes of individuals in the population, depending on their allelic state at the selected locus. Then, in a second step, we use techniques of the structured coalescent to reconstruct the genealogy of a population sample conditioned on this frequency path.

Consider, once again, a single locus under selection with two alleles A and a with mutation at rate u from a to A and back-mutation from A to a at rate v. The fitness values of A and a are $1+s$ and 1, respectively. Assume that we know the frequency x_t of allele A for all times t. Consider first a single individual of type A at some generation t. The parent of this individual in the previous generation could have been an A individual or a mutated a individual. Let δ be the timespan of one generation, such that $x_{t-\delta}$ is the frequency of A individuals in the previous generation. Then the proportion of A types in generation t that come form an A-type parent in the previous generation relative to new A-mutants (with a-type parent) is $x_{t-\delta}(1-v)(1+s) : u(1-x_{t-\delta})$. We measure time in units of $2N$ generations, such that $\delta = 1/(2N)$, and define $\Theta_u = 4Nu$, $\Theta_v = 4Nv$, and $\alpha = 2Ns$. In the coalescent scaling, we let $N \to \infty$, or $\delta \to 0$, and $s, u, v \to 0$, such that $\Theta_u, \Theta_v, \alpha = \text{const}$. We then obtain a backward mutation rate of

$$p_u(t) = \lim_{\delta\to 0} \frac{\Theta_u\big(1-x_{t-\delta}\big)}{2x_{t-\delta}\big(1-\Theta_v\delta/2\big)\big(1+\alpha\delta\big)+\big(\Theta_u\delta/2\big)\big(1-x_{t-\delta}\big)} = \frac{\Theta_u\big(1-x_t\big)}{2x_t}\,, \tag{2.13}$$

where we assume that the path x_t is continuous in this limit. For the Wright-Fisher model or the Moran model this follows from diffusion theory. Mutation form A to a is analogous. In addition, we need to account for coalescence events either among A individuals or among a individuals. If the numbers of A and a individuals in our sample is n_A and n_a, respectively, we obtain the following rates for four types of events,

$$p_{\text{coal},A}(t) = \binom{n_A}{2}\frac{1}{x_t}\ ;\quad p_{\text{coal},a}(t) = \binom{n_a}{2}\frac{1}{1-x_t} \tag{2.14}$$

$$p_u(t) = \frac{n_A\,\Theta_u\big(1-x_t\big)}{2x_t}\ ;\quad p_v(t) = \frac{n_a\,\Theta_v\,x_t}{2\big(1-x_t\big)}\,. \tag{2.15}$$

We make the following observations:

- The conditioned coalescent with a biallelic locus under selection corresponds to a structured coalescent with two islands, where the population sizes of the islands are allowed to change in time. Here, $2Nx_t$ and $2N(1-x_t)$ are the sizes of the "A island" and the "a island", respectively. Mutation corresponds to migration, where Θ_u and Θ_v are the forward rates and p_u and p_v the backward rates.

- Selection does not appear explicitly in the equations (2.14). As long as selection is sufficiently weak such that terms $\sim su$ and $\sim s/N$ can be ignored (which is implicit to the scaling limit), it does not affect the conditioned mutation or coalescence rates. This holds even if selection changes with time and/or with the allele frequency, i.e., $s = s(t, x)$. Of course, this does not mean that selection has no effect on the genealogies: the effect of selection is included through its impact on the allele trajectory x_t.

- As always, we can add neutral mutation to the coalescent genealogy. Since neutral mutation does not interfere with descent, this can be done by *mutation dropping* in a last step. Neutral mutations occur at a constant rate $\Theta_n/2$ per line, leading to a Poisson distributed number of mutations with parameter $L_b\Theta_n/2$ for a branch of length L_b.

- Clearly, we can extend the framework to more than two alleles at the selected locus. It is also straightforward to add real spatial structure, as long as the allele frequencies across all demes are known. With d islands and k alleles, we then obtain a structured model with $d \cdot k$ classes.

So far, we have assumed that we directly investigate the genealogy of the selected locus (i.e., the selected site itself and a region that is tightly liked to it). In praxis, this is often not true. More frequently, we have a stretch of sequence from a neutral locus in the neighborhood of the selected locus, with a recombination distance of r between both loci. It is easy to include recombination into the conditioned coalescent framework and thus to describe the genealogy of a linked neutral locus. We define a rescaled recombination rate $\rho = 2Nr$, which is kept constant after letting $N \to \infty$. As always, this has the effect that several events (recombination, mutation, and coalescence) do not happen simultaneously. Recombination occurs at a constant rate ρ per line, but only a fraction of these events matters for the genealogy: we only need to consider events where the genetic background at the selected locus changes from A to a or vice-versa. Specifically, consider an individual at time t that is associated with the A allele. Following the genealogy at the neutral locus one generation back in time, there are two possibilities (ignoring new mutation): either our focal individual derives from an A-parent that has not just recombined with an a individual, or it derives form an a-parent that has recombined with an A individual. The relative frequencies for these cases are $x_{t-\delta}\big(1-r(1-x_{t-\delta})\big)$ for the former and $rx_{t-\delta}\big(1-x_{t-\delta}\big)$ for the latter. We thus obtain a backward recombination rate for A individuals to change to the a background of (in units of $2N$ generations)

$$p_{\text{reco},A}(t) = \lim_{\delta\to 0} \frac{2Nrx_{t-\delta}\big(1-x_{t-\delta}\big)}{x_{t-\delta}\big(1-r(1-x_{t-\delta})\big)+rx_{t-\delta}\big(1-x_{t-\delta}\big)} = \rho\,\big(1-x_t\big)\,.$$

For n_A A alleles and n_a a alleles in a sample, we thus obtain

$$p_{\text{reco},A}(t) = n_A\rho\,\big(1-x_t\big)\;;\quad p_{\text{reco},a}(t) = n_a\rho\,x_t\,, \tag{2.16}$$

which complement the rates for mutation and coalescence described above. Like mutation at the selected locus, also recombination has the effect of migrating from one "island" (genetic background) to the other.

To obtain explicit results, we now need to make assumptions about the trajectory of the selected allele x_t. In principle, the most appropriate framework to model allele trajectories at a single locus is diffusion theory, which uses the same scaling assumptions as the coalescent. In praxis, however, this is difficult: under the diffusion, x_t is not a simple function, but a stochastic path. We thus need to average over all these paths to obtain a distribution of coalescent genealogies. There are two ways to proceed. Numerically, we can rely on forward-time computer simulations of a finite Wright-Fisher population and average over the resulting paths. Analytically, further progress is usually only possible in the deterministic limit where x_t is a unique function of time. We will consider the most important examples below.

Balancing selection

The easiest scenario is the case of frequency-dependent selection, where balancing selective forces keep the frequency of the A allele at an intermediate value. This is what happens, for example, in a diploid population with overdominance at the selected locus, or due to any mechanism that leads to a fitness advantage of the rare type (e.g. if predators specialize on the more frequent type). If we ignore stochastic fluctuations in the deterministic limit, we simply have $x_t = \bar{x} = \text{const}$. In terms of the structured coalescent, we thus have a scenario with two islands of constant size and constant, but generally asymmetric backward migration rates (where M is twice the rate following the convention),

$$M_{A\to a} = (1-\bar{x})\Big(2\rho+\frac{\Theta_u}{\bar{x}}\Big)\quad;\quad M_{a\to A} = \bar{x}\Big(2\rho+\frac{\Theta_v}{1-\bar{x}}\Big)\,. \tag{2.17}$$

Among other things, we can now calculate the expected coalescence time and heterozygosity for a sample of size two taken from the linked neutral locus. We need to distinguish three different states: both lines associated with allele a, both lines associated with A, and one line each associated with a and A. For the heterozygosity, in particular, we obtain the following recursions ($\Theta_n = 4Nu_n$ is the neutral mutation rate),

$$H_{AA} = \frac{\Theta_n}{\Theta_n + M_{A\to a} + (1/\bar{x})} + \frac{M_{A\to a}\, H_{Aa}}{\Theta_n + M_{A\to a} + (1/\bar{x})}\,, \tag{2.18}$$

$$H_{Aa} = \frac{\Theta_n}{\Theta_n + (M_{A\to a} + M_{a\to A})/2} + \frac{M_{A\to a}\, H_{aa} + M_{a\to A}\, H_{AA}}{2\Theta_n + M_{A\to a} + M_{a\to A}}\,, \tag{2.19}$$

$$H_{aa} = \frac{\Theta_n}{\Theta_n + M_{a\to A} + 1/(1-\bar{x})} + \frac{M_{a\to A}\, H_{Aa}}{\Theta_n + M_{a\to A} + 1/(1-\bar{x})}\,. \tag{2.20}$$

We obtain, for example,

$$H_{Aa} = \frac{2\Theta_n + M_{A\to a}\,\frac{\Theta_n + M_{a\to A}\, H_{Aa}}{\Theta_n + M_{a\to A} + 1/(1-\bar{x})} + M_{a\to A}\,\frac{\Theta_n + M_{A\to a}\, H_{Aa}}{\Theta_n + M_{A\to a} + (1/\bar{x})}}{2\Theta_n + M_{A\to a} + M_{a\to A}}\,, \tag{2.21}$$

and thus

$$H_{Aa} = \frac{2\Theta_n + \frac{\Theta_n M_{A\to a}}{\Theta_n + M_{a\to A} + 1/(1-\bar{x})} + \frac{\Theta_n M_{a\to A}}{\Theta_n + M_{A\to a} + (1/\bar{x})}}{2\Theta_n + \frac{M_{A\to a}(\Theta_n + 1/(1-\bar{x}))}{\Theta_n + M_{a\to A} + 1/(1-\bar{x})} + \frac{M_{a\to A}(\Theta_n + (1/\bar{x}))}{\Theta_n + M_{A\to a} + (1/\bar{x})}}\,. \tag{2.22}$$

For $\bar{x} = 1/2$ and $\Theta_u = \Theta_v = \Theta$, we have two islands of equal size and symmetric migration with $M = \rho + \Theta$ and

$$H_{Aa} = \Theta_n\,\frac{\Theta_n + 2\Theta + 2\rho + 2}{\Theta_n(\Theta_n + 2\Theta + 2\rho + 2) + 2\Theta + 2\rho}\,, \tag{2.23}$$

in accordance with the result (1.57) of the symmetric island model. Usually, $\Theta << 1$, but ρ varies depending on the distance to the selected locus. We thus obtain the coalescent of a strongly structured population in the direct neighborhood of the selected site, but only a weak effect (corresponding to the strong migration limit) at a larger distance.

Background selection

As a second application, we consider a genealogy that is affected by recurrent deleterious mutation in the genomic background. At first, we assume that there is no recombination (such as in clonal reproduction). Each deleterious mutation occurs at a new site in the genome and we assume that all mutations have the same effect s. The fitness of a genotype then depends solely on the number of deleterious mutations it carries. We also assume that mutational effects are independent. The fitness of the jth type then is $w_j = 1 - js$. Mutation (by one step from type j to type $j+1$) occurs at a constant rate of u for all types. Let f_j be the frequency of the jth type in the population. We can then show

Lemma *The Poisson*

$$f_j = \exp[-u/s]\frac{(u/s)^j}{j!} \tag{2.24}$$

is the stationary distribution of the process.

Proof We have the following system of differential equations

$$\dot{f}_0 = (1 - \bar{w}) f_0 - u f_0 \,, \tag{2.25}$$

$$\dot{f}_j = (w_j - \bar{w}) f_j + u(f_{j-1} - f_j) \quad j \geq 1 \,. \tag{2.26}$$

In equilibrium ($\dot{f}_j = 0$), we immediately obtain the mean fitness $\bar{w} = 1 - u$ and

$$f_j = \frac{u}{js} f_{j-1} = \frac{(u/s)^j}{j!} f_0 = \frac{(u/s)^j}{j!} \Big(\sum_{k=0}^{\infty} \frac{(u/s)^k}{k!} \Big)^{-1} = \frac{(u/s)^j}{j!} \exp[-(u/s)] \,, \tag{2.27}$$

confirming the Poisson claim. It is easy to extend this derivation to a general epistatic fitness function $w(j) = w_j$.

- The classical approach is to address the problem in discrete time. Here, fitness is measured on a direct scale and the effects of single mutations are multiplicative. This is, the fitness of the jth type ($j = 0, 1, 2, 3, \dots$) is $w_j = (1 - s)^j$. We assume that the number of new mutations is Poisson distributed with parameter u. This is, the probability of k new mutations is

  $$m_k = \exp[-u] \frac{u^k}{k!} \tag{2.28}$$

  We can then again show that the equilibrium distribution is Poisson as given by (2.24). Note first that the mean fitness of the population is

  $$\begin{aligned} \bar{w} &= \sum_j f_j w_j = \sum_j \exp[-u/s] \frac{(u/s)^j}{j!} (1 - s)^j \\ &= \exp[-u] \sum_j \exp[-((u/s) - u)] \frac{((u/s) - u)^j}{j!} = \exp[-u] \,. \end{aligned} \tag{2.29}$$

  With selection before new mutation, we then have

  $$\begin{aligned} f'_j &= \sum_{k=0}^{j} f_k \frac{w_k}{\bar{w}} m_{j-k} = \sum_{k=0}^{j} \exp[-u/s] \frac{(u/s)^k}{k!} \frac{(1 - s)^k}{\exp[-u]} \exp[-u] \frac{u^{j-k}}{(j - k)!} \\ &= \exp[-u/s] \frac{1}{j!} \sum_{k=0}^{j} \frac{j!}{k!(j - k)!} ((u/s) - u)^k u^{j-k} = \exp[-u/s] \frac{(u/s)^j}{j!} = f_j \,. \end{aligned} \tag{2.30}$$

From the equilibrium distribution, we immediately obtain the backward mutation rates for the coalescent. For continuous time, the backward rate from class j to class $j - 1$ becomes

$$2Nu \frac{f_{j-1}}{f_j} = 2Nu \frac{j}{u/s} = \alpha \cdot j \,, \tag{2.31}$$

with $\alpha = 2Ns$. If we consider the process backward in time, we obviously have the zero-mutation class as absorbing state. We first focus on the genealogy of a single individual. We can derive the average time that is needed for an individual currently in class j to reach absorption in class 0

$$\mathrm{E}[T_{j\to 0}] = \sum_{k=1}^{j} \frac{1}{k\alpha} \approx \frac{\log[j+1]}{\alpha}\,. \tag{2.32}$$

(The calculation for discrete time is a bit more complex, but yields similar results.) For a sample of size two, we can estimate the probability that coalescence occurs before both lines have reached the zero class. For this, assume that both individuals start out with j mutations. The probability for coalescence in this class before either line migrates to class $j-1$ is

$$\frac{1/f_j}{2\alpha j + 1/f_j} = \frac{1}{2\alpha j \cdot f_j + 1} = \frac{1}{2\alpha j \cdot \frac{(\Theta/2\alpha)^j}{j!}\exp[-(\Theta/2\alpha)] + 1}$$

This value is very small (and thus coalescence can be ignored) for all classes with $f_j \gg 1/(2\alpha j)$. On the other hand, all classes with $f_j \ll 1$ can be safely ignored in the model anyway. We thus see that for moderately strong selection of $\alpha > 10$, coalescence is very unlikely before the lines have reached class 0. In other words: the mutations carried by both individuals very likely have different origin. (This holds, at least, if $\Theta < 4\alpha$: for very strong mutation, lines typically carry many mutations and can coalesce in a low mutation class, such as f_1, before f_0 is reached). We conclude the following:

- For moderately strong selection $\alpha > 10$ and $\Theta \leq 4\alpha$, the background selection model is governed by "strong migration". All lines migrate back to the class $j = 0$, where the genealogy is given by the standard neutral coalescent with a reduced effective population size of $N_e = N\exp[-\Theta/2\alpha]$ (i.e., the size of the $j = 0$ class).

- For weak selection, $\alpha < 10$, mutations are rare (and background selection has no effect) if mutation is even weaker, $\Theta < \alpha$. However, if mutation is not weak, $\Theta > \alpha$, background selection changes the coalescent histories in non-trivial ways.

Adding recombination To add recombination, we (first) make the simplifying assumption that all background selection occurs at a single locus at a recombination distance r from the neutral locus that is considered. At a recombination event, the genetic background at the selected locus is changed to a random background that is drawn from the (equilibrium) population. In the backward direction, this induces migration rates of the form ($\rho = 2Nr$)

$$p_{\mathrm{reco},i\to j} = \rho f_j \tag{2.33}$$

independently of the class prior to the recombination event. The effect is that the march to class zero is interrupted and restarted at rate ρ. With mutation and recombination, but without coalescence, we obtain the following system of differential equations in the backward direction

$$\dot{x}_0 = \alpha x_1 + \rho(f_0 - x_0)\,, \tag{2.34}$$
$$\dot{x}_j = -\alpha j x_j + \alpha(j+1)x_{j+1} + \rho(f_j - x_j),\,. \tag{2.35}$$

We see that there is no absorbing state anymore for $\rho > 0$. We obtain the equilibrium distribution from the conditions $\dot{x}_j = 0$ leading to

$$x_1 = \frac{\rho}{\alpha}\,(x_0 - f_0)\,, \tag{2.36}$$
$$x_j = \frac{j-1}{j}\,x_{j-1} + \frac{\rho}{j\alpha}\,(x_{j-1} - f_{j-1}) = \frac{j-1+\rho/\alpha}{j}\,x_{j-1} - \frac{\rho}{j\alpha} f_{j-1}\,, \tag{2.37}$$

which can, in principle, be solved (resulting in a geometric mixture of Poisson distributions, see Durrett). We restrict our treatment to the case of small u/s, where we can ignore all classes beyond $j = 1$. We then have

$$x_0 + x_1 = x_0 + \frac{\rho}{\alpha}(x_0 - f_0) = 1 \quad \Rightarrow \quad x_0 = \frac{1 + (\rho/\alpha) f_0}{1 + \rho/\alpha} \tag{2.38}$$

and

$$x_1 = \frac{\rho}{\alpha}\frac{1 - f_0}{1 + (\rho/\alpha)} \approx \frac{\Theta\rho}{2\alpha(\alpha+\rho)}\,. \tag{2.39}$$

To include coalescence, we can argue as follows: in the presence of recombination, the migration rates of the system are increased. We can thus safely assume the strong migration limit in all cases where this limit is valid even for $\rho = 0$. The coalescence rate then follows as

$$\begin{aligned} p_{\text{coal}} &= \frac{x_0^2}{f_0} + \frac{x_1^2}{f_1} \approx \frac{\left(1 - \frac{\Theta\rho}{2\alpha(\alpha+\rho)}\right)^2}{1 - (\Theta/2\alpha)} + \frac{\left(\frac{\Theta\rho}{2\alpha(\alpha+\rho)}\right)^2}{\Theta/2\alpha}\,. \\ &\approx \frac{\big(2\alpha(\alpha+\rho) - \Theta\rho\big)^2(1 + (\Theta/2\alpha))}{(2\alpha(\alpha+\rho))^2} + \frac{2\alpha\Theta\rho^2}{(2\alpha(\alpha+\rho))^2} \\ &\approx 1 + \frac{-2(\alpha+\rho)\Theta\rho + (\alpha+\rho)^2\Theta + \Theta\rho^2}{2\alpha(\alpha+\rho)^2} = 1 + \frac{\Theta/2\alpha}{(1+\rho/\alpha)^2}\,. \end{aligned} \tag{2.40}$$

The corresponding expected time is

$$\mathrm{E}[T_{\text{coal}}] = \frac{1}{p_{\text{coal}}} \approx 1 - \frac{\Theta/2\alpha}{(1+\rho/\alpha)^2}\,. \tag{2.41}$$

As expected, we get back to the result without background selection for $\rho \to \infty$.

Background selection with multiple selected loci

As further extension of the background selection model, we consider that deleterious mutation will occur at many loci along a recombining chromosome. We assume that mutation happens at ℓ loci A_1 to A_ℓ with arbitrary mutation rates u_1 to u_ℓ and independent deleterious effects s_1 to s_ℓ.

Lemma *The distribution of the number of mutations at all loci A_i is independent Poisson with parameter u_i/s_i,*

$$f_{n_1,\ldots,n_\ell} = \prod_{i=1}^{\ell} f_{n_i} = \prod_{i=1}^{\ell} \frac{(u_i/s)^{n_i}}{n_i!} \exp[-(u_i/s_i)]\,, \tag{2.42}$$

where f_{n_i} is the marginal distribution at locus A_i.

Proof Note first that the assumption of independence of loci A_i in equilibrium implies that the distribution is invariant under arbitrary modes of recombination among the loci. Since mutation, selection, and recombination act independently in the continuous time model, it is sufficient to show that the distribution is also invariant under mutation and selection. We have

$$\dot{f}_{0,\ldots,0} = (1-\bar{w})f_{0,\ldots,0} - \Big(\sum_i u_i\Big) f_{0,\ldots,0}$$

and thus $\bar{w} = \sum_i u_i =: u$. In linkage equilibrium, we can further consider the marginal dynamics of the allele frequencies at locus i,

$$\dot{f}_{n_i} = \big(\bar{w}_{n_i} - \bar{w}\big) f_{n_i} + u_i\big(f_{n_i-1} - f_{n_i}\big)\,.$$

With the marginal fitness, $\bar{w}_{n_i} = 1 - n_i s_i - \sum_{j\neq i} u_j$, and $\bar{w}_{n_i} - \bar{w} = u_i - n_i s_i$, this reduces to the dynamics in the single locus case and we can thus infer a Poisson equilibrium distribution at A_i with parameter u_i/s_i.

Backward process We now consider a neutral locus B that is linked to the selected loci with a recombination rate of ρ_i between B and A_i. The first essential insight is:

Lemma *Let X_{n_i} be the random variable of the marginal backward process that records only the number of deleterious mutations (the mutation class) at locus A_i. Then the backward dynamics for X_{n_i} is independent of the other selected loci A_j, $j \neq i$, and follows the model with a single selected locus, with u, r, and s replaced by u_i, r_i, and s_i (or Θ, ρ, α replaced by Θ_i, ρ_i, α_i).*

Proof Since mutation and recombination are independent, we can consider them separately. Backward mutation at locus A_i occurs at a rate

$$p_{\text{mut}}[(\dots, n_i, \dots) \to (\dots, n_i - 1, \dots)] = 2Nu_i \frac{f_{\dots,n_i-1,\dots}}{f_{\dots,n_i,\dots}} = 2Nu_i \frac{j}{u_i/s_i} = n_i\alpha_i$$

and is thus independent of the state at all other loci. Backward recombination replaces the genotype of all loci beyond the crossover point by a randomly drawn genome from the equilibrium distribution of the forward process. As we have seen above, this equilibrium has product structure and the distribution at locus A_i is again independent of the other loci. Note that this holds for any linkage structure and the position of B on the chromosome. Also multiple crossing over or gene-conversion events are possible.

- We can thus describe the marginal dynamics for all X_{n_i} separately. Since u_i correspond only to the mutation rate at a single locus, the assumption of small $u_i/s_i (= \Theta_i/2\alpha_i)$ is usually justified. As above, we can thus focus on states with zero or one mutation at each locus, $n_i \in \{0, 1\}$, with equilibrium probability

$$x_{n_i=1} = 1 - x_{n_i=0} = \frac{(\Theta_i/2\alpha_i)(\rho_i/\alpha_i)}{(1 + \rho_i/\alpha_i)}\,.$$

- We approximate the distribution of $X_{n_1,\dots,n_l}$ by a product of the marginal distributions

$$x_{n_1,\dots,n_\ell} = \prod_{i=1}^{\ell} x_{n_i}\,.$$

This is appropriate if genotypes rarely carry multiple deleterious mutations at strongly linked selected loci (u_i/s_i small and /or recombination between neighboring selected loci sufficiently large). In the strong migration limit, the coalescence rate then follows as

$$p_{\text{coal}} = \sum_{n_1,\dots,n_\ell=0,1} \prod_{i=1}^{\ell} \frac{x_{n_i}^2}{f_{n_i}} = \prod_{i=1}^{\ell} \Big(\frac{x_{n_i=0}^2}{f_{n_i=0}} + \frac{x_{n_i=1}^2}{f_{n_i=1}}\Big) \tag{2.43}$$

$$\approx \prod_{i=1}^{\ell} \Big(1 + \frac{\Theta_i/2\alpha_i}{(1+\rho_i/\alpha_i)^2}\Big) \approx \exp\Big[\sum_{i=1}^{\ell} \frac{\Theta_i/2\alpha_i}{(1+\rho_i/\alpha_i)^2}\Big]\,. \tag{2.44}$$

Continuous chromosome We can assume that deleterious mutation occurs continuously along the chromosome. For this, we define a *mutation density* $\theta(x)$, such that the mutation rate for a locus that extends from genome position a to b is given by $\int_a^b \theta(x)dx$. We also define a local selection intensity $\alpha(x)$. To capture recombination, we introduce the so-called *map position* $M(x)$ for each genome position x (where x is measured on the physical scale per *base pair*). $M(x)$ measures distance on a recombination scale (in *Morgans*), such that the recombination rate per time unit between genome positions x and y is $|M(x) - M(y)|$ (for a constant recombination density ρ, in particular, $|M(x) - M(y)| = \rho|x - y|$). Let L be the total length of the chromosome and let y be the position of the neutral locus. Dissecting the genome in ever smaller units (= loci), Eq. (2.44) then turns into

$$p_{\text{coal}}(y) = \exp\Big[\int_0^L \frac{\theta(x)\alpha(x)}{2(\alpha(x) + |M(x) - M(y)|)^2}\, dx\Big]\,. \tag{2.45}$$

From modern data sets, we usually have information about the nucleotide diversity $\pi(x)$ (heterozygosity at the nucleotide level) along the genome. In the infinite sites limit, $\pi(x)$ is proportional to the expected coalescent time of a sequence pair, and thus

$$\frac{\pi(y)}{\pi_0(y)} = \mathrm{E}[T_{\text{coal}}(y)] = \frac{1}{p_{\text{coal}}(y)} = \exp\Big[-\int_0^L \frac{\theta(x)\alpha(x)}{2(\alpha(x) + |M(x) - M(y)|)^2}\, dx\Big]\,. \tag{2.46}$$

where $\pi_0(y)$ is the neutral nucleotide diversity without background selection. In the standard neutral model, we have $\pi_0(y) = \Theta_n(y)$ in the infinite sites model, where $\Theta_n(y)$ is the neutral mutation rate in the region (including demography and population structure, $\pi_0(y)$ is still proportional to $\Theta_n(y)$).

Assume now that $\theta(x)$, $\alpha(x)$ and $M(x)$ can be approximated by smooth (and differentiable) functions. $\theta(x)$ and $\alpha(x)$ are bounded and $M(x)$ is monotonically increasing with x. Then the integral in Eq. (2.46) is dominated by an interval $[y - W, y + W]$ around y. Assume that the mutation density and the selection strength can be approximated by constant values in this interval, denoted as θ_y and α_y. Assume further that we can approximate $M(x) \approx M(y) + \rho_y(x - y)$ in this region. Then

$$\begin{aligned} \frac{\pi(y)}{\pi_0(y)} &\approx \exp\Big[-\int_{y-W}^{y+W} \frac{\theta_y\alpha_y}{2(\alpha_y + \rho_y|x - y|)^2}\, dx\Big] = \exp\Big[-2\int_0^W \frac{\theta_y\alpha_y}{2(\alpha_y + \rho_y x)^2}\, dx\Big] \\ &= \exp\Big[\frac{\theta_y}{\rho_y}\Big(\frac{\alpha_y}{\alpha_y + \rho_y W} - 1\Big)\Big] = \exp\Big[\frac{-\theta_y W}{\alpha_y + \rho_y W}\Big] \approx \exp\Big[\frac{-\theta_y}{\rho_y}\Big]\,. \end{aligned} \tag{2.47}$$

- As expected, strong recombination (large ρ_y) reduces the local effect of background selection and thus increases $\pi(x)$ towards π_0. We thus expect a positive correlation of $\pi(y)$ and ρ_y under background selection. This is indeed seen in many data sets. In particular, nucleotide diversity is reduced in regions of low recombination, such as around centromeres.

- We also see that the selection strength plays only a minor role and drops out for a sufficiently large window size W. We obtain

$$\pi(y) \sim \Theta_n(y) \cdot \exp\Big[\frac{-\theta_y}{\rho_y}\Big]\,,$$

  A positive correlation of $\pi(y)$ and ρ_y could thus also result if ρ is positively correlated with Θ_n (e.g., recombination is mutagenic) or if ρ is negatively correlated with the deleterious mutation density θ (e.g., the gene density is higher in low recombining regions). Both these confounding factors need to be excluded before background selection can be inferred.

Muller's ratchet

In our treatment of background selection, we have used a deterministic theory to derive the equilibrium allele frequencies for the genotype classes A_i. In particular, the Poisson distribution predicts that the frequency of the mutation-free class is

$$f_0 = \exp[-(u/s)] \, .$$

In a recombining population, the equilibrium distribution is a product over independent loci and the relevant mutation rate is that of a gene locus (or recombinational unit). With $u \approx 10^{-4}$ (per gene) and a typical $s \approx 0.002$, we have a frequency of $\exp[-0.5] \approx 0.6$ for the fittest class at each locus. Importantly, wildtype alleles won't get lost in a large finite population, despite of unidirectional mutation. In a large genome, genotypes will (almost) necessarily have deleterious mutations at some positions, but fitter combinations can always be recovered by appropriate recombination and are not lost.

In contrast, for a non-recombining population, the relevant unit (locus) is the whole genome. With $u \approx 0.1$, we have a frequency of $\exp[-50] \approx 2 \cdot 10^{-22}$. Even in a huge population, this class will almost never be represented in a population: it is quickly lost due to genetic drift; and without recombination and back mutation it cannot be recovered anymore. As it turn out, this has drastic consequences for the dynamics. If the class with zero mutations vanishes, genomes with a single mutation form a new "fittest class". If we set up a deterministic system with $f_0 = 0$ as initial condition, we obtain exactly same dynamics, but with the roles of the classes shifted by one and with a new mean fitness $\bar{w} = 1 - u - s$. This means, however, that now f_1 takes the "quasi equilibrium" value that we had determined for f_0 before. Hence, also the one-mutant class gets lost and the process is iterated, lowering the mean fitness with each step. Once the mean fitness is so low that the population cannot sustain itself anymore ($\bar{w} < 0$ in continuous time), it will die out.

This is the famous process of Muller's ratchet that describes an extinction risk of non-recombining populations. The ratchet can move only in one direction and "clicks" every time a fitness class has died out. There is considerable theory to describe the click-rate of the ratchet and to determine factors that can stop the process. This is generally difficult (and partially unsolved) and beyond the scope of this lecture. Insights include:

- Even very low recombination rates are sufficient to stop the ratchet. Many asexual organisms (such as bacteria) have mechanisms that allow them to recombine at a low rate. This can be sufficient to guarantee their long-term survival.

- Other mechanisms to decrease or stop the rate of the ratchet are beneficial mutations and positive epistasis among deleterious mutations.

- Finally, the ratchet may be stopped if the effective deleterious mutation rate goes down during the process. A good example is the human Y-chromosome that effectively evolves without recombination. Although the human Y- and X-chromosomes have probably derived from a common ancestor chromosome, the Y-chromosome today only carries ~ 78 genes, compared to > 1000 on the X-chromosome. We conclude that most original genes on the Y chromosome have already been lost, probably due to a ratchet-like process. Today, the effective deleterious mutation rate on the Y-chromosome is already much reduced. Also, since weakly selected genes will be lost most easily, the average selection pressure is increased. Both factors contribute to a slowing of further degradation.

2.3 Selective Sweeps

In the previous sections, we have been concerned with the impact of balancing selection and purifying selection (recurrent deleterious mutation) on coalescent histories. We will now turn to the third mode: positive selection. The paradigmatic scenario is one of a new beneficial mutation that arises in the population at some time t_0, quickly increases in frequency and fixes in the population. In contrast to balancing selection and background selection, we thus do not consider an equilibrium, but a transient phenomenon.

Consider a haploid population of size $2N$ and a single locus under selection with two alleles a and A. The fitness values are 1 and $1+s$, respectively. Assume that a single new A mutant appears in the population at time $t_0 = 0$. Let $x(t)$ be the frequency of A in the population at time t. As in the previous cases, the simplest approach is to model selection as a deterministic process. In continuous time, $x(t)$ changes under selection according to the logistic differential equation

$$\dot{x}(t) = \alpha \cdot x(t) \big(1 - x(t)\big) \; ; \quad x(0) = x_0 \, , \tag{2.48}$$

where $\alpha = 2Ns$ and time is measured on a scale of $2N$ generations. This is solved by

$$x(t) = \frac{x_0}{x_0 + (1 - x_0)\exp[-\alpha t]} \, . \tag{2.49}$$

The resulting model for the genetic footprint of positive selection is also called the *logistic sweep model*. The time t_ϵ for a logistic sweep to reach frequency $x(t_\epsilon) = 1 - \epsilon$ from a small starting frequency $x_0 = \epsilon$ is

$$1 - \epsilon = \frac{\epsilon}{\epsilon + (1 - \epsilon)\exp[-\alpha \, t_\epsilon]} \quad \Rightarrow \quad t_\epsilon = \frac{2\log[(1/\epsilon) - 1]}{\alpha} \approx \frac{-2\log[\epsilon]}{\alpha} \, . \tag{2.50}$$

With a single new mutant, the canonical choice is $\epsilon = 1/(2N)$, but due to stochastic effects the fixation time will be somewhat shortened and a different choice gives more accurate results. We can argue as follows: The deterministic differential equation captures the average frequency change of the stochastic trajectory, given the current frequency, $\dot{x} = f(x)$. For any frequency at a distance from the boundaries at $x = 1$ and $x = 0$

this is a reasonable approximation. For very small x-values, however, we need to account for the fact that a stochastic reduction of the frequency may lead to a loss of the allele from the population. However, for our model of a sweeping allele, we discard these cases and only consider the frequency paths that reach fixation at $x = 1$. This conditioning on non-extinction leads to an effective acceleration of $\dot{x}$ relative to the deterministic model for small x. Similarly, for large x near one, a small stochastic fluctuation is sufficient to drive the allele to fixation. As a consequence, the stochastic process reaches fixation earlier (on average) than predicted by the neutral path. To include this effect into our model, note that the total fixation time of a neutral allele in a population of size $2N$ is $4N$ generations, corresponding to $t_{\text{fix},n} = 2$ on the coalescent scale. In more general, the average time to reach a frequency x_0 is $2x_0$. We thus see that the conditioned process for the neutral allele is faster than the one of the beneficial allele as predicted by the deterministic model if

$$\dot{x} = \alpha\, x(1-x) < \frac{1}{2}$$

which will (approximately) be the case for $x < 1/(2\alpha)$. Since the fixation process of a beneficial allele should always be as least as fast as the one of a neutral allele, we replace the logistic increase by a linear increase as in the neutral case for $x < 1/(2\alpha)$ and for $x > 1 - 1/(2\alpha)$. This results in

$$t_{\text{fix}} = \frac{2 + 2\log[2\alpha - 1]}{\alpha} = \frac{2\log[\alpha]}{\alpha} + \mathcal{O}\Big[\alpha^{-1}\Big]\,. \tag{2.51}$$

Hitchhiking Consider now a neutral locus that is linked to the selected locus with a recombination rate r per generation or $\rho = 2Nr$ per $2N$ generations. Assume that two alleles B and b segregate at the neutral locus. Let $p_{bA}(t)$ be the frequency of the b allele among haplotypes with the A allele at the selected locus, and $p_{ba}(t)$ the frequency of b on a-haplotypes. The total frequency of b is thus $p_b(t) = p_{bA}(t)x(t) + p_{ba}(t)\big(1 - x(t)\big)$. We then obtain the following differential equation for the conditioned frequencies (forward in time):

$$\dot{p}_{bA}(t) = \rho\big(1 - x(t)\big)\big(p_{ba}(t) - p_{bA}(t)\big)\,, \tag{2.52}$$

$$\dot{p}_{ba}(t) = \rho \cdot x(t)\big(p_{bA}(t) - p_{ba}(t)\big)\,. \tag{2.53}$$

We thus have

$$\frac{\partial}{\partial t}\big(p_{bA}(t) - p_{ba}(t)\big) = -\rho\big(p_{bA}(t) - p_{ba}(t)\big) \tag{2.54}$$

and hence

$$p_{bA}(t) - p_{ba}(t) = \big(p_{bA}(0) - p_{ba}(0)\big)\exp[-\rho t]\,. \tag{2.55}$$

With this we obtain the general solution

$$p_{bA}(t) = p_{bA}(0) + \rho\big(p_{ba}(0) - p_{bA}(0)\big)\int_0^t (1 - x(t'))\exp[-\rho t']\,dt' \tag{2.56}$$

$$p_{ba}(t) = p_{ba}(0) + \rho\big(p_{bA}(0) - p_{ba}(0)\big)\int_0^t x(t')\exp[-\rho t']\,dt' \tag{2.57}$$

In particular, with the choice $p_{bA}(0) = 0$ and $p_{ba}(0) = 1$ we get

$$P_\rho := p_{bA}(t_{\text{fix}}) = \rho \int_0^{t_{\text{fix}}} \big(1 - x(t)\big) \exp[-\rho t]\, dt\,. \tag{2.58}$$

By integrating in parts, and with $x(0) = 0$ and $x(t_{\text{fix}}) = 1$, we also have

$$P_\rho = 1 - \int_0^{t_{\text{fix}}} \dot{x}(t') \exp[-\rho t']\, dt' = 1 - \int_0^1 \exp[-\rho t(x)]\, dx\,. \tag{2.59}$$

For given $x(t)$ (logistic or combination linear/logistic, see above), or inverse function $t(x)$, the integral can be evaluated numerically.

Backward in time, P_ρ represents the probability that an individual in the A population at time t_{fix} comes from the a-population at time $t = 0$. Since at time t_{fix} all of the population is in the A part (or almost all of the population if we take t_{fix} as end of the logistic phase), this corresponds to the probability that a line of decent from an individual sampled at the time of fixation of the beneficial allele *escapes the sweep* by coalescing into the a background. Alternatively, the line of descent will run back to the founder individual of the beneficial mutation. Any allele B that is initially associated with the beneficial A and that is found at frequency p_{Ba} in the a background will *hitchhike* to a higher frequency of

$$1 - P_\rho + P_\rho\, p_{Ba} = p_{Ba} + (1 - P_\rho)(1 - p_{Ba})\,.$$

Although Eq. (2.58) and (2.59) are exact and explicit expressions for the escape probability, the integral is inconvenient to evaluate. Simpler approximate expression can be obtained if we consider the process in the backward direction, as expressed by the following Lemma.

Lemma *Consider a logistic selective sweep for a strongly selected allele A. Then the escape probability to leading order in the selection strength α reads*

$$P_\rho = 1 - \exp\left[\frac{-\rho \log[\alpha]}{\alpha}\right] + \mathcal{O}\left[\frac{\rho}{\alpha}\right]. \tag{2.60}$$

Proof Note first that backward in time, the differential equation for the probability q_A that a line of descent is in the A population reads

$$\dot{q}_A = -\rho(1 - x(t))q_A + \rho \cdot x(t)(1 - q_A) = \rho[x(t) - q_A]\,, \tag{2.61}$$

where

$$p_{\text{reco},A}(t) = \rho\big(1 - x(t)\big) \quad \text{and} \quad p_{\text{reco},a}(t) = \rho \cdot x(t) \tag{2.62}$$

are the backward recombination rates. Although we cannot solve Eq. (2.61) directly, we can easily derive the probability that a line in the A background will stay there all the time and never recombine to the a background. As complement, we obtain the probability that a line of descent will recombine at least once,

$$\begin{aligned} P_\rho^+ &= 1 - \exp\Big[-\rho \int_0^{t_{\text{fix}}} \big(1 - x(t)\big) dt \Big] = 1 - \exp[-\rho t_{\text{fix}}/2] \\ &= 1 - \exp\Big[\frac{-\rho \log[\alpha]}{\alpha}\Big] + \mathcal{O}\Big[\frac{\rho}{\alpha}\Big], \end{aligned} \tag{2.63}$$

where we use the symmetry $x(t) = 1 - x(t_{\text{fix}} - t)$ of the deterministic trajectory. Clearly, P_ρ^+ is an upper bound for the escape probability P_ρ. To estimate the accuracy of this approximation, we can estimate the probability of at least a double recombination event of a single line, from A to a and back to A,

$$P_\rho^{++} = 1 - \exp\Big[-\rho^2 \int_0^{t_{\text{fix}}} \int_0^{t_1} x(t_2)\big(1 - x(t_1)\big)\, dt_2\, dt_1 \Big]. \tag{2.64}$$

Let t_ϵ be the time when $x(t)$ reaches some small value ϵ with $1/(2\alpha) < \epsilon < 1/2$. Because of symmetry, we also have $x(t_{\text{fix}} - t_\epsilon) = 1 - \epsilon$. We can then split the double integral as follows,

$$\begin{aligned} &\rho^2 \int_0^{t_{\text{fix}}} \int_0^{t_1} x(t_2)\big(1 - x(t_1)\big)\, dt_2\, dt_1 \\ &= \rho^2 \Big[\int_{t_\epsilon}^{t_{\text{fix}} - t_\epsilon} \int_0^{t_1} + \int_0^{t_\epsilon} \int_0^{t_1} + \int_{t_{\text{fix}} - t_\epsilon}^{t_{\text{fix}}} \int_0^{t_1} \Big] x(t_2)\big(1 - x(t_1)\big)\, dt_2\, dt_1 \\ &\le \rho^2 \Big[\int_{t_\epsilon}^{t_{\text{fix}} - t_\epsilon} \int_0^{t_1} \frac{1}{4}\, dt_2\, dt_1 + \int_0^{t_\epsilon} \int_0^{t_1} \epsilon\, dt_2\, dt_1 + \int_{t_{\text{fix}} - t_\epsilon}^{t_{\text{fix}}} \int_0^{t_1} \epsilon\, dt_2\, dt_1 \Big] \\ &= \frac{\rho^2}{8} t_{\text{fix}}\big(t_{\text{fix}} - 2t_\epsilon\big) + \rho^2 \epsilon t_\epsilon t_{\text{fix}} \le \frac{\rho^2}{2\alpha^2} \log\big[2\alpha - 1\big] \log\big[(1/\epsilon) - 1\big] + \frac{\rho^2\, 2\epsilon (\log[2\alpha - 1])^2}{\alpha^2}. \end{aligned}$$

since $t_{\text{fix}} - 2t_\epsilon = 2\log[(1/\epsilon) - 1]/\alpha$ is the time for the process to increase form ϵ to $1 - \epsilon$ and $t_\epsilon \le t_{\text{fix}}/2$. With the choice $\epsilon = 1/\log[\alpha]$, we thus have

$$P_\rho^{++} \le \frac{\rho^2\big(\log[2\alpha]\big)^2}{\alpha^2} \left(\frac{\log\big[\log[\alpha]\big]}{2\log[2\alpha]} + \frac{2}{\log[\alpha]} + \mathcal{O}\Big[\log[\alpha]^{-1}\Big] \right). \tag{2.65}$$

We see that P_ρ^{++} is of lower order than P_ρ^+ for large α. To leading order, we can thus can ignore multiple recombination events during the sweep and thus have $P_\rho \sim P_\rho^+$, proving the Lemma.

- We need to have $(\rho \log[\alpha])/\alpha$ of the order unity to obtain a non-trivial P_ρ^+ (i.e., $0 \ll P_\rho^+ \ll 1$). The Lemma shows that double recombinations can be ignored in this case if α is sufficiently large. From the proof we see that the factor that controls double recombinations scales like $\log[\log[\alpha]]/\log[\alpha]$ and is thus very weak. This leads to the question how good the approximation is for practical applications

(α of $10^2 - 10^4$). However, direct comparison of the approximation with the exact result Eq. (2.58) shows that it generally performs quite well. Note also that the approximation improves for a larger sample. For a sample of size n, $(\rho \log[\alpha])/\alpha$ of order $1/n$ is sufficient to see lines that escape the sweep. The probability for double recombination in at least one line is then suppressed by another factor $1/n$.

Next, consider a pair of individuals sampled at fixation of the beneficial allele. We are interested in the coalescence probability of the corresponding lines of descent during the selective sweep. Our central result is the following Theorem.

Theorem: Star-like approximation *To leading order for strong selection, the probability that two lines of descent do* not *coalesce during the time t_{fix} of the selective sweep, but to go back to different ancestors at time t_0 is*

$$p_{22} = 1 - (1 - P_\rho)^2 + \mathcal{O}\Big[\frac{\rho}{\alpha}\Big] = P_\rho(2 - P_\rho) + \mathcal{O}\Big[\frac{\rho}{\alpha}\Big] \approx 1 - \exp\Big[\frac{-2\rho\log[\alpha]}{\alpha}\Big]. \tag{2.66}$$

Proof If both lines go back to the founder of the beneficial mutation, an event with probability $(1 - P_\rho)^2$, they will certainly coalesce within the time t_{fix}. If only one line goes back to the origin of the beneficial allele, an event with probability $2P_\rho(1 - P_\rho)$, both lines clearly do not coalesce. Finally, both lines escape with probability P_ρ^2. In this case, they can either coalesce or not coalesce. To prove the Theorem, we need to show that the coalescence probability in this case is of lower order for large α and therefore does not affect Eq. (2.66). Remember that the coalescence rates at time t in the A and a population are

$$p_{\text{coal},A} = \frac{1}{x(t)} \quad ; \quad p_{\text{coal},a} = \frac{1}{1 - x(t)}. \tag{2.67}$$

We now need to distinguish two cases:

1. Either both lines recombine from A to a independently and then coalesce in the a background. The probability of this event can be estimated as

$$\begin{aligned} P_{rc} &\le 2\rho^2 \int_0^{t_{\text{fix}}} \int_0^{t_1} \int_0^{t_2} \frac{\big(1 - x(t_1)\big)\big(1 - x(t_2)\big)}{1 - x(t_3)}\, dt_3\, dt_2\, dt_1 &(2.68)\\ &\le 2\rho^2 \int_0^{t_{\text{fix}}} \int_0^{t_1} \int_0^{t_2} \big(1 - x(t_1)\big)\, dt_3\, dt_2\, dt_1 = \rho^2 \int_0^{t_{\text{fix}}} t_1^2 \big(1 - x(t_1)\big) dt_1 &(2.69)\\ &\le \frac{\rho^2 t_{\text{fix}}^3}{6} \approx \frac{4\rho^2 \log^3[\alpha]}{3\alpha^3} &(2.70)\end{aligned}$$

   In the region of interest with a non-trivial P_ρ this is smaller than P_ρ by a factor of $\log[\alpha]/\alpha$ and can thus be ignored for large α.

2. The alternative is that the two lines coalesce first in the A-domain and recombine to the a background later. This case is a bit more complex since we need to take

the initial linear increase of the trajectory due to genetic drift, $x(t) = t/2$, explicitly into account. The logistic part sets in at $t_0 = 1/\alpha$, where the trajectory has reached frequency $x_0 = 1/(2\alpha)$. We obtain

$$\begin{aligned} P_{cr} &\leq \rho \int_0^{t_{\text{fix}}} \int_0^{t_1} \frac{\big(1 - x(t_2)\big)}{x(t_1)}\, dt_2\, dt_1 = \rho \bigg[\int_0^{t_0} \int_0^{t_1} + \int_{t_0}^{t_{\text{fix}}} \int_0^{t_1} \bigg] \frac{\big(1 - x(t_2)\big)}{x(t_1)}\, dt_2\, dt_1 \\ &\leq \rho \bigg[\int_0^{t_0} \int_0^{t_1} \frac{1}{t_1/2}\, dt_2\, dt_1 + \int_{t_0}^{t_{\text{fix}}} \int_0^{t_1} \Big(1 + 2\alpha \exp\big[-\alpha(t_1 - t_0) \big] \Big)\, dt_2\, dt_1 \bigg] \\ &\leq 2\rho t_0 + \frac{\rho t_{\text{fix}}^2}{2} + \frac{2\rho}{\alpha}\big(1 + \alpha t_0\big) \leq \frac{6\rho}{\alpha} + \mathcal{O}\Big[\frac{\rho(\log[\alpha])^2}{\alpha^2} \Big] . \end{aligned} \tag{2.71}$$

In the relevant parameter range, with P_ρ between 0 and 1, this probability is smaller than P_ρ by a factor $\sim 1/(\log[\alpha])$. We thus see that this term is also small for large α, proving the Theorem.

- We can easily extend the reasoning of the Theorem to a sample of size n: each line goes back to the origin of the beneficial mutation with probability $1 - P_\rho = \epsilon^{\rho/\alpha}$ and escapes with probability P_ρ. All lines with the founder of the beneficial mutation as ancestor coalesce, and to leading order in α all escape lines will not coalesce. The probability that k out of n lines remain after the sweep is thus

$$p_{nk} \approx \binom{n}{k-1} P_\rho^{k-1} (1 - P_\rho)^{n-k+1} \quad \text{for } k < n\,, \tag{2.72}$$

$$p_{nn} \approx P_\rho^n + n P_\rho^{n-1}(1 - P_\rho) \quad \text{for } k = n\,. \tag{2.73}$$

This scheme, where the fate of individual lines (i.e., escape the sweep or not) is independent is also called the star-like approximation.

- For two lines that are caught in the sweep we can calculate the distribution for the coalescence time and the frequency of the A allele at this time. The probability that both lines coalesce before time $t_1 > t_0 = 1/\alpha$ is

$$\begin{aligned} \exp\Big[-\int_{t_1}^{t_{\text{fix}}} \frac{1}{x(t)}\, dt \Big] &\approx \exp\Big[-\int_{t_1}^{t_{\text{fix}}} \big(1 + 2\alpha \exp[-\alpha(t - t_0)] \big)\, dt \Big] \\ &= \exp\Big[t_1 - t_{\text{fix}} - 2\exp\big[-\alpha(t_1 - t_0) \big] + 2\exp\big[-\alpha(t_{\text{fix}} - t_0) \big] \Big] \\ &\approx \exp\Big[-2\exp\big[-\alpha(t_1 - t_0) \big] + \mathcal{O}\big[\log[\alpha]/\alpha \big] \Big] . \end{aligned}$$

For $t_1 = t_0$, we see that there is a probability of $\exp[-2] \approx 0.135$ that they coalesce before time t_0, i.e., in the short initial drift-phase of the selective sweep with frequency of the A allele $x < 1/(2\alpha)$. We also find that there is a 90% chance that coalescence happens before time $t_1 = 4/\alpha$, corresponding to a frequency of $x < 10/\alpha$. Analogous results hold for larger samples. For very strong selection, all coalescence thus happens almost directly at the point of origin of the beneficial allele, giving rise to a star-like shape of the genealogy. For more realistic values, such as $\alpha = 1000$, we see that coalescence occurs typically at low frequencies ($x < 1\%$) and relatively early ($4/(2\log[2\alpha]+2) \approx 0.23$, i.e., with probability 0.9 in the first 23% of the total sweep time), but not exactly star-like.

- The factor in Eq. (2.71) suppressing escape from A to a of a pair of lines that has already coalesced in the A domain is only logarithmic in α. The problem gets worse with larger sample sizes: Since the coalescence probability scales with n^2, but recombination only with n, the probability for escape of coalesced lines relative to single-line escape increases proportionally to n. As a consequence, the star-like approximation, Eqs. (2.66) and (2.72), although true for $\alpha \to \infty$, is only of modest quality for typical selection strengths, in particular if the sample size is large. However, they are usually sufficient to demonstrate the qualitative effects of a selective sweep. Improved analytical approximations based on a stochastic sweep model are available (e.g. Durett & Schweinsberg; Etheridge, Pfaffelhuber, Wakolbinger), but problems with large samples remain.

Assume now that we sample two individuals τ generations after fixation of the beneficial allele. Consider the nucleotide diversity π at a neutral locus linked to the selected locus with recombination distance ρ. We can prove the following

Theorem *In the infinite sites model, we have*

$$\mathrm{E}[\pi] = \Theta\Big(1 - (1-p_{22})\exp[-\tau]\Big)\,. \tag{2.74}$$

Proof Note first that for the infinite-sites model $\mathrm{E}[\pi] = \Theta \cdot \mathrm{E}[T_s]$, where T_s is the pair-coalescence time in the presence of the sweep. Let $\mathrm{E}[T_0]$ be the expected coalescence time without a sweep. The genealogies with and without a sweep will be different if and only if (i) coalescence does not occur before time τ and (ii) the two lines coalesce during the sweep. (i) is an event with probability $\exp[-\tau]$ and (ii) is an independent event of probability $(1-p_{22})$. Further, if coalescence has not happened until time τ, the remaining expected time to coalescence in the case without sweep is $\mathrm{E}[T_0|T_0>\tau]-\tau = 1$ because of the memoryless property of the Markov process and $\mathrm{E}[T_0] = 1$. We thus have

$$\mathrm{E}[T_s] = 1-\big(\mathrm{E}[T_0]-\mathrm{E}[T_s]\big) = 1-(1-p_{22})\exp[-\tau]\big(\mathrm{E}[T_0|T_0>\tau]-\tau\big) = 1-(1-p_{22})\exp[-\tau]$$

which proves the claim.

Recap: Estimators for Θ and text statistics

Most tests based on the site frequency spectrum use a common principle: they compare two estimators of the population mutation parameter Θ, that should be equal under the standard neutral model. Significant deviations lead to rejection of standard neutrality.

The scenarios of population demography and selection lead to typical patterns for these deviations.

Estimators Let S_i be the number of polymorphic sites of size i in the sample. Under standard neutrality, we can define an unbiased estimator $\hat{\Theta}_i$ for each size class,

$$\mathrm{E}[S_i] = \frac{\Theta}{i} \quad \longrightarrow \quad \hat{\Theta}_i := i \cdot S_i \,. \tag{2.75}$$

Widely used estimators are linear combinations of the $\hat{\Theta}_i$. They can be distinguished according to the relative weight that is put on a certain class. The most important ones are the following:

1. *Watterson's estimator*,

$$\hat{\Theta}_{\mathrm{W}} := \frac{S}{a_n} = \frac{1}{a_n}\sum_{i=1}^{n-1} S_i = \frac{1}{a_n}\sum_{1\le i\le n/2} \tilde{S}_i \,, \tag{2.76}$$

uses the total number of segregating sites and puts an equal weight on each mutation class. The last equation expresses $\hat{\Theta}_{\mathrm{W}}$ in terms of frequencies of the folded spectrum. The distribution of $\hat{\Theta}_{\mathrm{W}}$ is independent of coalescent topologies, but only depends on the coalescent times.

2. Let π_{ij} be the number of differences among two sequences i and j from our sample. We have $\mathrm{E}[\pi_{ij}] = \mathrm{E}[S(n = 2)] = \Theta$. Averaging over all pairs, this leads to the *diversity-based estimator* (or *Tajima's estimator*),

$$\hat{\Theta}_\pi := \frac{2}{n(n-1)}\sum_{i<j} \pi_{ij} \,. \tag{2.77}$$

We can also express $\hat{\Theta}_\pi$ in terms of the (folded) frequency spectrum as follows,

$$\hat{\Theta}_\pi = \binom{n}{2}^{-1}\sum_{i=1}^{n-1} i(n-i)S_i = \binom{n}{2}^{-1}\sum_{1\le i\le n/2} i(n-i)\tilde{S}_i \,. \tag{2.78}$$

$\hat{\Theta}_\pi$ puts the highest weight on classes with an intermediate frequency. It also depends on the distribution of tree topologies.

3. *Fay and Wu's estimator*,

$$\hat{\Theta}_{\mathrm{H}} := \binom{n}{2}^{-1}\sum_{i=1}^{n-1} i^2 S_i \,, \tag{2.79}$$

puts a hight weight on mutation classes of the unfolded spectrum with a high frequency of the derived allele. It is not a summary statistic of the folded spectrum, but requires knowledge of the ancestral state.

4. The *singleton estimator* $\hat{\Theta}_{\mathrm{s}}$ uses the singletons of the folded spectrum,

$$\hat{\Theta}_{\mathrm{s}} := \frac{n-1}{n}\Big(S_1 + S_{n-1}\Big) = \frac{n-1}{n}\tilde{S}_1\,. \tag{2.80}$$

   It has all its weight at both ends of the unfolded spectrum.

Test statistics

1. *Tajima's D*,

$$D_{\mathrm{T}} := \frac{\hat{\Theta}_\pi - \hat{\Theta}_{\mathrm{W}}}{\sqrt{\mathrm{Var}[\hat{\Theta}_\pi - \hat{\Theta}_{\mathrm{W}}]}}\,. \tag{2.81}$$

   D_{T} is negative if we have an excess of very low or very high frequency alleles, whereas it is positive if many sites segregate at intermediate frequencies.

2. *Fu and Li's D*,

$$D_{\mathrm{FL}} := \frac{\hat{\Theta}_{\mathrm{W}} - \hat{\Theta}_{\mathrm{s}}}{\sqrt{\mathrm{Var}[\hat{\Theta}_{\mathrm{W}} - \hat{\Theta}_{\mathrm{s}}]}} \tag{2.82}$$

   D_{FL} is an extreme version of Tajima's D. It focuses entirely on the signal from singletons of the folded spectrum. It is more powerful for the type of pattern that it is constructed for (excess of singletons), but also less general and more vulnerable to sequencing errors. The power also decreases with larger samples.

3. *Fay and Wu's H*,

$$H_{\mathrm{FW}} := \frac{\hat{\Theta}_\pi - \hat{\Theta}_{\mathrm{H}}}{\sqrt{\mathrm{Var}[\hat{\Theta}_\pi - \hat{\Theta}_{\mathrm{H}}]}}\,. \tag{2.83}$$

   H_{FW} focuses on high-frequency derived alleles. In contrast to D_{T} and D_{FL}, it weighs singletons in the opposite direction to high-frequency derived alleles. The main idea (and advantage) of H_{FW} is that it distinguishes positive selection from population growth. A (potential) problem is that it requires reliable estimates of the ancestral state from one or several outgroups.

The footprint of positive selection

We have now sufficient information to get an overview of the "average" footprint of recent positive selection. Consider a sample of size n taken from a neutral locus at recombination distance ρ from the selected site directly at the time of fixation of the beneficial allele ($\tau = 0$ in Eq. 2.74). We assume that the time of the sweep is sufficiently short that neutral mutation during this time can be ignored. Let $\mathrm{E}_0[\Theta_\pi]$ be the expected nucleotide diversity without a sweep. We then have

$$\frac{\mathrm{E}[\Theta_\pi]}{\mathrm{E}_0[\Theta_\pi]} = p_{22} = 1 - (1-P_\rho)^2 = 1 - \exp\Big[-2\rho\log[\alpha]/\alpha\Big] \tag{2.84}$$

- We have $\mathrm{E}[\Theta_\pi]/\mathrm{E}_0[\Theta_\pi] = 0$ for $\rho = 0$, and a linear increase with slope $2\log[\alpha]/\alpha$ for small ρ.

- The width of the sweep region (on a recombination scale) scales like $\sim \alpha/\log[\alpha]$. In more general, this scale is given by the inverse of the fixation time. For $\rho = 0.5\alpha/\log[\alpha]$, we are back to 63.2% of the background diversity, for $\rho = \alpha/\log[\alpha]$ at 86.5%, and for $\rho = 2\alpha/\log[\alpha]$ at 98.2%. For $\alpha = 1000$ we have $\alpha/\log[\alpha] \approx 145$. For fruitflies, $r \approx 10^{-8}$ per base pair and $2N \approx 10^6$; we thus have a width of the sweep region extending over 14.5 kb. This should be compared to typical gene lengths of $1 - 10\,kb$.

- Note that footprints of single selective sweeps do not look like the "average" pattern. In particular, there is a broad central region without any recombination lines. The width of this central region is exponentially distributed with mean (and standard deviation)
$$\frac{\alpha}{n\log[\alpha]}\,.$$
To each side of the center, the median distance to the first recombination event is
$$\big(1 - P_\rho\big)^n = \frac{1}{2} \quad \Rightarrow \quad \rho = \alpha\log[2]/\big(2n\log[\alpha]\big)$$
A typical size of this central region is thus several kb. On both sides, the central region is followed by a region with a single escape line. These width of these flanking regions is also exponentially distributed: in the star-like approximation, the corresponding mean/median width is the same as the central region, with the number of lines n reduced by one.

For the expected number of segregating sites, we get from the simple star-like approximation:
$$\mathrm{E}[\Theta_W] = \sum_{k=2}^{n} p_{nk}\Theta\frac{a_k}{a_n} \tag{2.85}$$
where $a_k = \sum_{i=1}^{k-1}(1/i)$ and once again mutation during the sweep is ignored.

- For small ρ, $\mathrm{E}[\Theta_W]$ increases like
$$\frac{\mathrm{E}[\Theta_W]}{\mathrm{E}_0[\Theta_W]} = \frac{\mathrm{E}[\Theta_W]}{\Theta} \approx p_{n2}\frac{a_2}{a_n} = \frac{n\log[\alpha]}{a_n\alpha}\rho + \mathcal{O}\big[\rho^2\big]\,.$$
For $n > 2$, the increase is thus faster than the one of π: we thus expect to have $\pi - \Theta_W < 0$. For $n = 10$ ($n = 100$) we get a recovery of: 70.6% (82.2%) for $\rho = \alpha/(2\log[\alpha])$; 87.3% (91.3%) for $\rho = \alpha/\log[\alpha]$; 97.6% (97.4%) for $\rho = 2\alpha/\log[\alpha]$.

- Of course, the central region without variation and $\Theta_W = \Theta_\pi = 0$ is the same one as derived above. In the flanking regions with a single escape line, we have

$$\mathrm{E}[\Theta_W] = \frac{\Theta}{a_n} \left[= \frac{\Theta}{2.83} \text{ for } n = 10 \right] \quad ; \quad \mathrm{E}[\Theta_\pi] = \frac{2\Theta}{n} \left[= \frac{\Theta}{5} \text{ for } n = 10 \right] . \tag{2.86}$$

  We thus again see a much faster initial increase to the standard neutral expectation of Θ_W relative to Θ_π. In more general, if there are $1 < k < n$ lines left after the sweep,

$$\begin{aligned} \mathrm{E}[\Theta_W] &= \frac{\Theta\, a_k}{a_n} , \\ \mathrm{E}[\Theta_\pi] &= \frac{\big((k-1)(n-k+1) + (k-1)(k-2)/2\big)\Theta}{n(n-1)/2} = \frac{(k-1)(2n-k)\Theta}{n(n-1)} . \end{aligned}$$

  (For $\mathrm{E}[\Theta_\pi]$, we sum over two component: pairwise differences among the $k-1$ lines that have recombined out of the sweep and the differences of these $k-1$ lines to each of the $n-k+1$ lines that coalesce in the sweep.)

- Finally, we can derive expected number of mutations of size i for the star-like logistic sweep model (once again without new mutations after the sweep). Assume that k lines survive the sweep. The expected frequency of mutations of size i *before the sweep* is $E'[S_i] = \Theta/i$ for $1 \leq i \leq k-1$. After the sweep, we need to account for the $n-k+1$ lines that coalesce in the sweep: all mutations that sit on the ancestor of these lines will appear $n-k+1$ times in our sample. For a given mutation of size i before the sweep, there is thus a chance of i/k that it will appear as a mutation of size $n-k+i$ after the sweep. With probability $(k-i)/k$ it will still be seen as a mutation of size i. Summarizing we have, conditional on k lines surviving the sweep,

$$E[S_i] = \begin{cases} \frac{\Theta}{i} - \frac{\Theta}{k} & i \leq \min[k-1, n-k] \\ \frac{\Theta(k-i)}{k \cdot i} + \frac{\Theta}{k} = \frac{\Theta}{i} & n-k+1 \leq i \leq k-1 \\ 0 & k \leq i \leq n-k \\ \frac{\Theta}{k} & i \geq \max[k, n-k+1] . \end{cases} \tag{2.87}$$

  For the folded spectrum, where mutations with size i and $n-i$ are collected in the same class, we obtain

$$E[\tilde{S}_i] = \begin{cases} \frac{\Theta}{i} & i \leq \min[k-1, n-k] \\ \frac{\Theta}{i} + \frac{\Theta}{n-i} & n-k+1 \leq i \leq \frac{n}{2} \\ 0 & k \leq i \leq \frac{n}{2} . \end{cases} \tag{2.88}$$

  The spectrum clearly shows the deficit of intermediate frequency polymorphism for $k \leq n/2$. The unfolded spectrum also shows a surplus of high-frequency derived alleles relative the standard neutral case, in particular for small k. By averaging over k with the weights $p_{nk} = p_{nk}(\rho)$, we can also obtain the expected spectrum as a function of the recombination distance to the selected site:

$$E[S_i|\rho] = \sum_{k=i+1}^{n} p_{nk}(\rho)\Big(\frac{\Theta}{i} - \frac{\Theta}{k}\Big) + \sum_{k=n-i+1}^{n} p_{nk}(\rho)\frac{\Theta}{k}\,. \tag{2.89}$$

- Above we have assumed that the frequency spectrum in the absence of a selective sweep is the standard neutral one. An alternative is that we do not make this assumption, but start with an empirical spectrum that has been measured from genome-wide data. This is the approach that is taken in the *sweepfinder* software (Nielsen *et al.* 2005). Assume that the frequency of mutations of size i in a sample of size n is q_i, $i < n$ (this is the information we have from data). Then the frequency of mutations of size j among the k lines that still exist at the start of the sweep is

  $$q_{j,k} = \sum_{i=j}^{n-1} q_i \,\frac{\binom{i}{j}\binom{n-i}{k-j}}{\binom{n}{k}}\,. \tag{2.90}$$

  Then the normalized frequency spectrum after a sweep follows as

  $$\frac{E[S_i|\rho]}{\sum_{i=1}^{n-1} E[S_i|\rho]} = \sum_{k=i+1}^{n} p_{nk}(\rho) q_{i,k}\,\frac{k-i}{k} + \sum_{k=n-i+1}^{n} p_{nk}(\rho) q_{i-n+k,k}\,\frac{i-n+k}{k}\,. \tag{2.91}$$

  With $q_{i,k} = 1/i$ for all k we reproduce formula (2.89) above (up to normalization) [cf formula 6 in Nielsen *et al.*, with $k \to k+1$ due to slightly different definition of the p_{nk}].

2.4 Soft selective sweeps

So far, we have assumed that positive selection acts on a single copy of a new beneficial mutation: adaptation is mutation limited. In many cases, however, multiple copies of the later-beneficial allele may already be present in the population when the selection pressure sets in. In this case, the population can adapt form this so-called *standing genetic variation.* Alternatively, the beneficial allele can also arise in the population multiple time by recurrent mutation (or also by immigration) during the sweep phase. Both processes are not captured in the basic sweep model and lead to deviating selective footprints called *soft selective sweeps* (in contrast to the classical *hard sweeps*).

Adaptation from standing genetic variation In the context of the star-like approximation, we have seen that lineages that coalesce during a (hard) selective sweep will typically do so at very low frequency of the beneficial allele. If adaptation occurs from standing genetic variation (SGV), this means that coalescence will likely happen prior to the environmental change at time t_0 unless the allele frequency x_0 at this time is even smaller. If we assume that each of the $2Nx_0$ copies in the SGV either establishes or gets lost with an independent probability $2s_b$ (where s_b is the selection coefficient), we can employ a Poisson distribution to obtain

$$P_{\mathrm{sgv}}(x_0) \approx 1 - \exp[-4Nx_0 s_b] \tag{2.92}$$

for the probability of successful adaptation from the SGV

$$P_{\mathrm{mult}}(x_0) \approx \frac{1 - (1 + 4Nx_0 s_b)\exp[-4Nx_0 s_b]}{1 - \exp[-4Nx_0 s_b]} \tag{2.93}$$

for the probability that more than a single allele from the SGV contributes to the adaptation, conditioned on that successful adaptation occurs in the first place. If the later-beneficial allele segregates in the population in mutation-selection-drift balance prior to the environmental change, we need to integrate these expressions over corresponding stationary allele frequency distribution. For previously neutral derived allele, we have

$$f(x_0) \approx \Theta\, x_0^{\Theta - 1}$$

resulting in

$$P_{\mathrm{sgv}} \approx \Theta \log[4Ns_b] \qquad ; \qquad P_{\mathrm{mult}} \approx 1 - \frac{1}{\log[4Ns_b]}$$

in the limit of small Θ. We see that, for neutral SGV, *conditioned on* that a sweep from the SGV occurs at all this sweep will typically lead to the fixation of multiple copies.

The footprint of a sweep from the SGV depends on the allele frequency trajectory x_t of the beneficial allele, for neutral SGV, x_t changes only due to drift prior to the environmental change. In a simple model, we can assume that x_t is constant during this time. We can then describe the genealogy of all lines that are still caught in the sweep at time t_0 by an island model with migration rates (backward in time),

$$p_{\mathrm{reco},A} = n_A \rho \left(1 - x_0\right) \; ; \quad p_{\mathrm{reco},a} = n_a \rho\, x_0 \, ,$$

see Eq. (2.16). Since the A island is small, we can ignore recombination (migration) back to this island, $p_{\mathrm{reco},a} \approx 0$. The dynamics is then equivalent to the "scattering phase" of the infinite-islands model, with migration replaced by recombination. In contrast to the star-like model, the rate of recombination relative to coalescence does not depend on time. There is thus no particular time that would favor one over the other more than any other time: both occur in parallel. This implies that groups of coalesced lineages can recombine and escape the A island just like single lineages. As in the infinite island model, the distribution of the number and size of these groups is given by the Ewens sampling formula (1.43). The net effect of the "standing phase" of the genealogy is thus that recombination lines can also produce intermediate-frequency polymorphism for a soft sweep from SGV. The pattern of the sweep in summary statistics like $\hat{\Theta}_\pi$, $\hat{\Theta}_W$, or Tajima's D_T is therefore less pronounced than for a hard sweep.

Adaptation from recurrent new mutation So far, we have assumed that the beneficial sweep allele has a single origin by mutation, either before or after the onset of positive selection. However, with recurrent mutation, it can also originate multiple times. We have already discussed the effect of mutation at the selected locus above in the case of balancing selection. In the genealogy, recurrent mutation is a second way (in addition to recombination) how lineages can cross from one genomic background to the other. If Θ_u is the mutation parameter to describe mutation to the beneficial allele and Θ_v is the parameter of back migration, we have

$$p_u(t) = \frac{n_A\,\Theta_u\big(1-x_t\big)}{2x_t} \qquad ; \qquad p_v(t) = \frac{n_a\,\Theta_v\,x_t}{2\big(1-x_t\big)}\,.$$

In the case of a selective sweep, all lineages are initially on the A island. We see that the rate of crossing over to the ancestral background due to recurrent beneficial mutation is $\sim (1-x_t)/x_t$ and therefore increases strongly as x_t becomes small – just like coalescence events, $p_{\text{coal}}(t) = \binom{n_A}{2}/x_t$. For a fast sweep (strong selection), both recurrent mutation and coalescence events before x_t is small. In this case case, we can also ignore the probability that a lineage will ever migrate back (via back mutation) onto the A island once it has emigrated. Consider the genealogy of a sample of size n and assume that recombination can be ignored (i.e., very close to the selected site or in a clonal population). Initially, all lines are associated with the A allele and we have two events, coalescence and emigration due to recurrent mutation. In the relevant region for small x_t the probability that the next event is mutation is

$$\frac{p_u(t)}{p_u(t)+p_{\text{coal}}(t)} = \frac{\Theta_u}{\Theta_u+n-1}\,. \tag{2.94}$$

This ratio is independent of the allele trajectory x_t and therefore does not depend on time nor on the selection strength (which enters through the shape of x_t). In more general, recurrent beneficial mutation leads to multiple haplotypes form the pre-sweep population in a pot-sweep sample (i.e., a soft sweep) if any recurrent mutation event occurs before the most recent common ancestor is reached. The probability for a soft sweep therefore follows as

$$P_{\text{soft}}(n) = 1 - \prod_{i=1}^{n-1} \frac{i}{\Theta_u+i} \approx \Theta_u \log[n]\,. \tag{2.95}$$

We see that soft sweeps from recurrent mutation become relevant when $\Theta_u = 4N_e u$ is of the order of 0.1. This is typically the case when either the effective population size N_e is large (e.g., fruitflies or microbial populations) or if the allelic mutation rate is large. The latter is the case, in particular, for adaptive loss-of-function mutations, when shutting down a gene function (that is e.g., exploited by a parasite) is beneficial. Obviously, there are many ways to disrupt a gene, leading to a high mutation rate u.

Since the sequence of coalescence and emigration events does not depend on x_t for a recurrent-mutation soft sweep, the number and sizes of groups of lines that escape the sweep is once again given by the Ewens sampling formula, just like in the case of escape by recombination at a SGV soft sweep. Therefore, also the consequences on the summary statistics of the site-frequency spectrum (such as $\hat{\Theta}_\pi$, $\hat{\Theta}_W$ or D_T) are also the same. In particular, recurrent-mutation soft sweeps produce more intermediate-frequency polymorphism than classical hard sweeps and are therefore more difficult to detect by tests that are based on allele frequencies. Still, the sweep pattern is not the same as in the SGV case. Haplotypes that are introduced by recombination are only visible in the flanking region of the selected locus (beyond the recombination breakpoint). In contrast haplotypes from recurrent mutation run right across the selected site. The produce a clear pattern that can be detected with haplotype tests that go beyond measures of allele frequencies at single sites.

3 Selection in structured populations

In all the lecture so far, our aim has been to describe neutral genetic variation and how it is shaped by either spatial population structure or by selection. As a tool, we have mostly used the structured coalescent process. However, we have never directly studied the frequencies of alleles under selection in a structured population. This is of importance, in particular, since selection coefficients can change among demes. We then need to know how selection and migration (or gene-flow) act together to determine the frequencies of selected alleles.

We will address this problem with the island model. However, in contrast to our treatment of neutral variation, we will choose a forward-time formalism and ignore genetic drift. This is a reasonable approximation for selected alleles if selection is much stronger than drift, i.e., $2Ns \gg 1$. We will always assume this in the following and set $N \to \infty$. For simplicity, we will also ignore new mutation and focus on the interaction of migration and selection. Again, this is often appropriate since mutation is usually a weak force. There is a large body of literature on island models with inhomogeneous selection, starting with Haldane and Wright. We will only discuss the most basic model, where selection acts on a single locus.

We consider a population that is distributed over d discrete demes and evolves in discrete generations. The life cycle starts at the zygote state. Selection on a single diploid locus with two alleles A and a acts in each deme prior to migration. Finally, random mating (within each deme) and reproduction produces a new generation of zygotes in Hardy-Weinberg equilibrium, separately in all demes. The fitness values for the genotypes AA, Aa, and aa in deme i are denoted as $w_i(AA)$, $w_i(Aa)$, and $w_i(aa)$, respectively. Migration is defined via an ergodic backward migration matrix $\mathbf{M}$ with entries m_{ij} giving the fraction of individuals (zygotes) in deme i with parents in deme j. In particular, $m_{ii} = 1 - \sum_j m_{ij}$. The migration-selection dynamics is then given by the following equation system

$$p_i' = \sum_j m_{ij}\, p_j\, \frac{w_j}{\bar{w}_j}\,, \quad 1 \le i \le d\,, \tag{3.1}$$

where w_j is the marginal fitness in deme j,

$$w_j = w_j(AA)\, p_j + w_j(Aa)\,(1 - p_j)\,,$$

and

$$\bar{w}_j = w_j(AA)p_j^2 + w_j(Aa)2p_j(1-p_j) + w_j(aa)(1-p_j)^2$$

is the local mean fitness. We see that the dynamics only depends on ratios of fitness values. We can thus set one fitness value per deme to 1 without restriction. In the following, we will usually choose $w_j(aA) = 1$.

3.1 Protected polymorphism

We are interested in the the long-term fate of the alleles A and a under the dynamics. Essential information about this long-term behavior is given by the equilibrium points $p_i' = p_i$ and their stability. As it turns out, however, we can only explicitly derive these equilibria for some particular cases. A somewhat simpler problem is the identification of conditions that guarantee the maintenance of a genetic polymorphism. A so called protected polymorphism results if the dynamics implies an increase of the allele frequency once this frequency becomes sufficiently low. Note that $\mathbf{p}_0 = (p_1, \dots, p_d) := (0, \dots, 0)$ (absence of the A allele from all demes) and $\mathbf{p}_1 := (1, \dots, 1)$ (fixation of A) are always equilibria of the dynamical system. Mathematically, instability of equilibrium $\mathbf{p}_0$ is a sufficient condition for maintenance of the$A$ allele. Analogously, maintenance of the a allele is guaranteed if $\mathbf{p}_1$ is unstable. We thus need to determine the stability of these two monomorphic (boundary) equilibria. We can express the Jacobian of (3.1) as

$$\mathbf{J} = \mathbf{MD}\,.$$

At $\mathbf{p}_0$, we have

$$D_{ij} = \frac{\partial[p_i w_i/\bar{w}_i])}{\partial p_j}\Big|_{\mathbf{p}_0} = \delta_{ij}\frac{w_i(aA)}{w_i(aa)} = \frac{\delta_{ij}}{w_i(aa)}\,.$$

Thus,

$$J_{ij} = \frac{m_{ij}}{w_j(aa)} > 0\,.$$

Since $\mathbf{M}$ is ergodic and $\mathbf{J}^k \geq (\mathbf{M}/\max_j[w_j(aa)])^k$, also the Jacobian matrix $\mathbf{J}$ is ergodic. Let $\lambda_{\max}$ be the maximal eigenvalue of J. According to the Perron-Frobenius theorem, it is always real and uniquely determined. Allele A is protected if $\lambda_{\max} > 1$ and it is not protected if $\lambda_{\max} < 1$ (for $\lambda_{\max} = 1$ higher-order terms matter). The maximal eigenvalue satisfies

$$\min_i \sum_j J_{ij} \leq \lambda_{\max} \leq \max_i \sum_j J_{ij}$$

with equality (on both sides) if and only if all rows are equal. We conclude the following:

- Assume that (aA) is at least as fit as (aa) in all demes and $w_i(aA) = 1 > w_i(aa)$ for at least one deme. Then
$$\min_i \sum_j J_{ij} \geq \min_i \sum_j m_{ij} = 1 \,.$$
Since at least one row sum of $\mathbf{J}$ is larger 1, we have $\lambda_{\max} > 1$ and A is protected.

- Similarly, A is not protected if (aa) is at least as fit as (Aa) in all demes and $w_i(aA) = 1 < w_i(aa)$ for at least one deme.

- Equivalent equations hold for the protection of the a allele if we replace $w_j(aa)$ by $w_j(AA)$ in the entries of $\mathbf{J}$. If both alleles are protected, there is a protected polymorphism. We see that a sufficient condition is that $1 = w_i(aA) \geq w_i(aa)$ and $w_i(aA) \geq w_i(AA)$ in all demes and $w_i(aA) < w(aa)$ and $w_j(aA) < w_j(AA)$ in at least one deme each (which could be the same). Note that this condition is only slightly weaker than requiring overdominance in all demes.

Weak migration

- A more relaxed condition can be derived for weak migration. For this, assume that $w_i(aa) < 1 = w_i(aA)$ for at least one deme, but arbitrary fitness values in the other demes. Then a new A allele in this deme will be able to increase in frequency (*invade* the population) if migration is sufficiently weak. Indeed, we have
$$\frac{p_i' - p_i}{p_i}\bigg|_{\mathbf{p}\to\mathbf{p}_0} = \sum_j m_{ij}\,\frac{p_j}{p_i}\,\frac{w_j}{\bar{w}_j}\bigg|_{\mathbf{p}\to\mathbf{p}_0} - 1 \;\geq\; m_{ii}\,\frac{w_i}{\bar{w}_i}\bigg|_{\mathbf{p}\to\mathbf{p}_0} - 1 = \frac{m_{ii} - w_i(aa)}{w_i(aa)} \,.$$
Equivalent arguments hold for a. We thus have a protected polymorphism of alleles A and a if demes i and j exist with
$$w_i(aa) < m_{ii} = 1 - \sum_{k\neq i} m_{ik} \quad ; \quad w_j(AA) < m_{jj} = 1 - \sum_{k\neq j} m_{jk} \,.$$
We also conclude that $w_i(aa) < 1$ and $w_j(AA) < 1$ for some i and j is a sufficient condition for a protected polymorphism in the limit of weak migration $m_{ij} \to 0$, $\forall i, j$. I.e., we only require that heterozygotes are superior to aa genotypes in at least one deme, and superior to AA genotypes in another deme (which can be the same).

3.2 Levene model

A special case of the island model with selection is the Levene model. Here, we assume that the individuals from all demes (or, equivalently, the gametes they produce) enter a common migrant pool. They mix and mate (form zygotes) in this pool before they are redistributed to the single demes where selection occurs. We thus have population structure only with respect to selection (local competition for resources), but not with respect to reproduction. When assessing the population in zygote state (before selection), this means that the proportion of individuals with ancestors from any given deme i is equal in all demes, $m_{ij} = m_j$. In the backward migration matrix $\mathbf{M}$, we thus have all rows equal. As a consequence, we have equal allele frequencies $p_i = p_j := p$ in all demes at this stage already after a single generation. The dynamical equation then reads

$$p' = \sum_j m_j \, p \, \frac{w_j}{\bar{w}_j} \,. \tag{3.2}$$

For the protection of the A allele, we obtain the condition

$$\left.\frac{p' - p}{p}\right|_{\mathbf{p}\to 0} = \sum_j \frac{m_j}{w_j(aa)} - 1 \,,$$

and equivalently for the protection of the a allele. We can express the condition for a protected polymorphism as condition for the harmonic mean fitness of migrants

$$\Big(\sum_j \frac{m_j}{w_j(aa)}\Big)^{-1} < 1 \quad , \quad \Big(\sum_j \frac{m_j}{w_j(AA)}\Big)^{-1} < 1 \,. \tag{3.3}$$

Usually, the m_j are taken proportional to the deme sizes in the Levene model. For equal deme sizes, in particular, $m_j = 1/d$ and we obtain a protected polymorphism for

$$\frac{d}{\sum_j \frac{1}{w_j(aa)}} < 1 \quad , \quad \frac{d}{\sum_j \frac{1}{w_j(AA)}} < 1 \,. \tag{3.4}$$

- Since the harmonic mean is smaller or equal than the arithmetic mean (strictly smaller if there is any inhomogeneity at all), we always have a protected polymorphism in the Levene model if the arithmetic mean fitness of both the A and the a allele is smaller or equal to one. An example is the symmetric model with two demes and $w_1(AA) = 1 + s = w_2(aa)$ and $w_1(aa) = 1 - s = w_2(AA)$.

- Demes with small fitness values contribute a high weight to the harmonic means. In particular, if there are demes with $w_i(AA) < 1/d$ and $w_j(aa) < 1/d$, this will already guarantee a protected polymorphism.

- In a model with multiplicative fitness, we can parametrize $w_i(aa) = w_i^2(a)$; $w_i(aA) = w_i(a)w_i(A)$, and $w_i(AA) = w_i^2(A)$. The protected polymorphism condition then reads

$$\frac{1}{\sum_j m_j \frac{w_j(A)}{w_j(a)}} < 1 \quad , \quad \frac{1}{\sum_j m_j \frac{w_j(a)}{w_j(A)}} < 1 \,. \tag{3.5}$$

  This is equivalent to the haploid model. The condition is fulfilled, in particular, if alleles a and A are equivalent, but favored in different demes.

- The protected polymorphism condition for the Levene model can be understood as a rare-type advantage of alleles under spatially heterogeneous selection when competition is local (also called "soft selection"). A rare allele will enjoy a large advantage in a deme where it is favored and only competes against inferior types. In contrast, the advantage of an individual with the frequent type in a deme where it is favored is small, since it mostly competes against its own type.

- For the Levene model with two alleles, one often chooses $w_i(aa) = 1 - s_i$ and $w_i(AA) = 1 - r_i$. The case of interest is the one where each allele is favored in one deme and disfavored in the other deme, say $s_1, r_2 < 0$, $r_1, s_2 > 0$. For general asymmetric migration with $m_2 = 1 - m_1$, this results in the condition for an protected polymorphism,
$$\frac{m_1}{s_2} + \frac{(1-m_1)}{s_1} < 1 \quad , \quad \frac{m_1}{r_2} + \frac{(1-m_1)}{r_1} < 1$$

Equilibria for the Levene model

From (3.2), the following equilibrium condition for the diploid Levene model reads $p = p'$ with

$$p' = \sum_j f_j(p) = \sum_j m_j \frac{p w_j}{\bar{w}_j} = \sum_j \frac{m_j (p^2 w_j(AA) + p(1-p) w_j(aA))}{p^2 w_j(AA) + 2p(1-p) w_j(aA) + (1-p)^2 w_j(aa)} \,.$$

For haploids or diploids with multiplicative fitness (no dominance) within demes, this simplifies to

$$p' = \sum_j \frac{m_j p w_j(A)}{p w_j(A) + (1-p) w_j(a)} \,.$$

For the general diploid case, we derive

$$\frac{\partial}{\partial p} f_j(p) = \frac{m_j}{\bar{w}_j^2} \Big(\big(2p w_j(AA) + (1-2p) w_j(aA)\big)\big(p^2 w_j(AA) + 2p(1-p) w_j(aA) + (1-p)^2 w_j(aa)\big) \\ - \big(p^2 w_j(AA) + p(1-p) w_j(aA)\big)\big(2p w_j(AA) + 2(1-2p) w_j(aA) - 2(1-p) w_j(aa)\big)\Big)$$

leading to

$$\frac{\partial}{\partial p} f_j(p) = \frac{m_j}{\bar{w}_j^2} \Big(p^2 w_j(AA) w_j(aA) + 2p(1-p) w_j(AA) w_j(aa) + (1-p)^2 w_j(aA) w_j(aa) \Big) > 0 \,.$$

Thus, $p' = \sum_j f_j(p)$ is monotonic in p. Results for discrete dynamical systems (see e.g. the Ecology lecture) show that a monotonic iteration function implies monotonic convergence of the dynamics to an equilibrium point. Oscillating convergence and complex dynamical behavior, such as limit cycles or chaos, are excluded. We thus see that the allele frequency p for the Levene model will always converge monotonically to an equilibrium point (which may depend on the starting condition). For an arbitrary number of demes, however, neither the equilibrium condition for the diploid model nor the one for the haploid case can be solved explicitly. For the diploid case, in particular, potential internal equilibria are roots of a polynomial of degree $2d-1$. It has indeed been shown that for biallelic Levene models with d demes any number of internal equilibria $\leq 2d-1$ can be produced for a suitable choice of fitness parameters. In the following we therefore restrict our analysis to the case of two demes, where explicit results are possible.

Two demes For the haploid Levene model (or the diploid model with multiplicative fitness) and $m_1 = m = 1 - m_2$, we obtain

$$p = p' = \frac{m\, p w_1(A)}{p w_1(A) + (1-p) w_1(a)} + \frac{(1-m)\, p w_2(A)}{p w_2(A) + (1-p) w_2(a)}\,, \tag{3.6}$$

leading to

$$\begin{aligned} p\Big(p^2 w_1(A) w_2(A) + p(1-p)\big(w_1(a) w_2(A) + w_1(A) w_2(a)\big) + (1-p)^2 w_1(a) w_2(a)\Big) \\ = p\Big(p w_1(A) w_2(A) + (1-p)\big(m w_1(A) w_2(a) + (1-m) w_2(A) w_1(a)\big)\Big) \end{aligned}$$

which results in $p = p_0 = 0$ or $p = p_1 = 1$ or

$$p = \hat{p} = \frac{1 - m\, w_1(A)/w_1(a) - (1-m) w_2(A)/w_2(a)}{(1 - w_1(A)/w_1(a))(1 - w_2(A)/w_2(a))}\,. \tag{3.7}$$

- If $w_i(A) > w_i(a)$ for both demes $i = 1, 2$ we have $p < 0$ in (3.7) and thus no intermediate equilibrium. Similarly, if $w_i(A) < w_i(a)$ for both demes, we have $p > 1$. Only the boundary equilibria with $p = 1$ and $p = 0$ exist in these cases.
- For $w_1(A) > w_1(a)$ and $w_2(A) < w_2(a)$ (or labels $1, 2$ interchanged), the denominator in (3.7) is always negative. We get $p > 0$ for a negative numerator, which holds exactly under the protection condition (3.5) for A. Similarly, one easily checks that the condition $p < 1$ exactly reproduces the protection condition of a (as it must because of symmetry). We thus find that we have a protected polymorphism if and only if an internal equilibrium exists.

For the diploid Levene model, the equilibria in addition to the monomorphic equilibria $p_0 = 0$ and $p_1 = 1$ are given by the roots of a third-order polynomial. With the normalization $w_{1,2}(Aa) = 1$, we obtain in the general case

$$\begin{aligned} \Big(-p w_1(AA) + (2p-1) + (1-p) w_1(aa)\Big)\Big(p^2 w_2(AA) + 2p(1-p) + (1-p)^2 w_2(aa)\Big) = \\ m\Big(p^2(w_2(AA) - w_1(AA)) + (1-p)^2(w_1(aa) - w_2(aa)) + p(1-p)(w_1(aa) w_2(AA) - w_2(aa) w_1(AA))\Big) \end{aligned} \tag{3.8}$$

For the symmetric model with $w_1 = w_1(AA) = w_2(aa)$, $w_2 = w_2(AA) = w_1(aa)$ and $m = 1/2$ this reduces to

$$2\Big(- pw_1 + (2p-1) + (1-p)w_2\Big)\Big(p^2 w_2 + 2p(1-p) + (1-p)^2 w_1\Big) = \\ (w_2 - w_1)\Big(p^2 + (1-p)^2 + p(1-p)(w_2+w_1)\Big) \quad (3.9)$$

with solutions $\hat{p} = 1/2$ and

$$p_\pm = \frac{1}{2} \pm \frac{\sqrt{4 + w_1^2 + w_2^2 - 6w_1w_2}}{2(w_1 + w_2 - 2)} \,.$$

A necessary condition for $0 < p_\pm < 1$ is that

$$(w_1 + w_2 - 2)^2 > 4 + w_1^2 + w_2^2 - 6w_1w_2 \quad \Leftrightarrow \quad 2w_1w_2 > w_1 + w_2 \,.$$

Comparing with (3.4) we see that this is just the condition that the polymorphism is *not* protected and both monomorphic equilibria are stable. In case of a protected polymorphism we thus have exactly one stable internal equilibrium at $\hat{p} = 1/2$ for the symmetric diploid model. In contrast, when the monomorphic equilibria are stable, cases with one or three internal equilibria exist. With three internal equilibria, the central equilibrium $\hat{p} = 1/2$ can be stable, maintaining the polymorphism even if rare alleles are not protected and the monomorphic equilibria are also stable.

3.3 The continent-island model

One of the simples scenarios of population structure is one of a single island that is connected to a continent from which it receives migrants. There is no migration in the opposite direction, either because migration is truly unidirectional (only downstream a river), or because the continental population is so much larger than the island population that back migration can be safely ignored. As a consequence, we can first solve the evolutionary dynamics for the continent separately. Usually, it is simply assumed that the continent settles for a monomorphic equilibrium. Interest now focuses on the dynamics on the island. A question that is often raised in this context is when a locally adaptive gene on the island can be maintained and when it is swamped by maladaptive gene-flow from the continent.

We consider a single diploid locus with two alleles. Allele A is locally adapted on the island, whereas the continental allele a is disfavored. We allow for dominance at the locus and define the fitnesses of the three genotypes AA, Aa, and aa as 1, $1 - hs$, and $1 - s$, respectively. Let p be the frequency of the island allele A. With discrete generations and census in the zygote state before selection and migration (i.e., Hardy-Weinberg proportions at census), the dynamical equation reads

$$p' = (1-m)\,p\,\frac{w_A}{\bar{w}} = (1-m)\,p\,\frac{1-(1-p)hs}{1-2p(1-p)hs-(1-p)^2 s} \,. \quad (3.10)$$

The equilibrium condition is always fulfilled for $p = 0$ (loss of the island type) and for

$$f(p) := \frac{(1-(1-p)hs)(1-m)}{1-s+2s(1-h)p+s(2h-1)p^2} - 1 = 0 \tag{3.11}$$

with solutions

$$p_\pm = \frac{-2s(1-h)+hs(1-m) \pm \sqrt{s^2h^2(1+m)^2+4ms(1-2h)}}{2s(2h-1)}\,. \tag{3.12}$$

We observe the following:

- A rare island allele will be able to invade the population if $p' - p = p \cdot f(p) > 0$ for $p \to 0$. This is the case if and only if

  $$f(0) = \frac{(1-hs)(1-m)}{1-s} - 1 > 0 \quad \Leftrightarrow \quad m < m_1 := \frac{s-hs}{1-hs}\,. \tag{3.13}$$

  For $m > m_1$, $p = 0$ (loss of the island type) is a locally stable equilibrium.

- For $m < m_1$, we have $f(0) > 0$ and $f(1) = -m < 0$, we thus have a protected polymorphism on the island. One can show that the population will indeed settle at a unique stable intermediate equilibrium.

- For $m > m_1$, both $f(0)$ and $f(1)$ are negative. We then have a stable equilibrium $p > 0$ if and only if both p_+ and p_- are in the interval $[0, 1]$. We now investigate when this is the case.

- In the interval $p \in [0, 1]$, $f(p)$ is an analytical function. For $m > m_1$, the equilibrium solutions $p_\pm$, where $f(p) = 0$, thus cannot cross the interval boundaries at $p = 0$ and $p = 1$ as we vary m. Since $f(p) < 0$ in the unit interval for $m \leq 1$, any solutions $p_\pm \in [0, 1]$ (for $m > m_1$) can only exist below some finite maximal migration rate $m_2 < 1$.

- Both p_+ and p_- are real if and only if the expression under the root in (3.12) is positive. We find that for $s > 2h - 1$ this is the case for any $m > 0$ (in particular, this holds for any $h \leq 1/2$). Since this is in contradiction to a finite maximal m_2, we can exclude this case (i.e. we can conclude that $p_\pm$ cannot be in the unit interval in this case).

- Assume now $0 < s < 2h-1$ and $h \leq 1$. We find that $f(p)$ is continuously differentiable for all p in this case and $f(p) < 0$ for $p \to \pm\infty$. We conclude that $f(p) \geq 0$ in $[p_-, p_+]$ and $f(p) < 0$ otherwise. From $f'(p) = 0$, we obtain exactly two extreme values of $f(p)$ (one maximum and one minimum) at

  $$\hat{p}_\pm = \frac{-(2h-1)(1-hs) \pm \sqrt{(2h-1)(2h-1-sh^2)}}{hs(2h-1)}\,.$$

Note that $\hat{p}_- < 0$. Finally, we have $f'(1) = -hs(1-m) < 0$ and

$$f'(0) = \frac{s(3h - 2 - s(2h^2 - h)}{(1-s)^2}\,(1-m)\,.$$

- For
$$0 < s < s_0 := \frac{3h-2}{2h^2-h} \tag{3.14}$$
we have $f'(0) > 0$ and thus a maximum of $f(p)$ in $[0,1]$ (and no other maximum can exist). Thus, both $p_\pm$ are in the interval $[0,1]$ if they exist. Note that the condition $s < s_0$ requires $h > 2/3$.
- For $s > s_0$, we have $f'(0) < 0$ and the number of extreme values between 0 and 1 must be even. Since $\hat{p}_- < 0$ it must be zero. This excludes $p_\pm \in [0,1]$.

- For $0 < s < 2h-1$ this requires
$$0 \leq m < m_2 := \frac{\sqrt{(2h-1-h^2s)(2h-1)} - (2h-1-h^2s)}{4h-2-2h^2s}\,. \tag{3.15}$$

- Finally, we also find $m_2 \geq m_1$ for $s < s_1$ and $m_2 = m_1$ for $s = 0$ and $s = s_0$. Summarizing, we have a stable internal equilibrium for
$$\begin{aligned} m &< m_2 \quad \text{for} \quad 0 \leq s \leq s_0\,,\\ m &< m_1 \quad \text{for} \quad s > s_0\,. \end{aligned} \tag{3.16}$$
For $h \leq 2/3$ (where $s_0 \leq 0$) always the bound m_1 applies and maintenance of the island allele implies that it can invade. This includes the case of $h < 0$ (overdominance). For $h \geq 1$ (i.e., underdominance) we have $m_1 \leq 0$ and the island allele can never invade. However, it can still be maintained if $m < m_2$. (For $h \leq 1$, all results apply as long as $hs < 1$, i.e., hybrids have positive fitness).

Continuous time model

The dynamics in continuous time is usually qualitatively equivalent to the discrete time model, but derivations are often easier. In continuous time, all evolutionary forces are described as rates and the corresponding events (selection, mutation, migrations, etc.) occur in parallel. For selection and migration in the continent-island model, we obtain

$$\dot{p} = p(w_A - \bar{w}) - mp = p\Big(-(1-p)hs + 2p(1-p)hs + (1-p)^2 s - m\Big) \tag{3.17}$$

Similarly to the discrete case, we define $f(p) = \dot{p}/p$,

$$f(p) = p^2(s - 2hs) + p(3hs - 2s) + s - hs - m\,, \tag{3.18}$$
$$f'(p) = p(2s - 4hs) + 3hs - 2s\,. \tag{3.19}$$

We observe:

- The island allele can invade if $f(0) > 0$, which implies
$$m < m_1 := s - hs\,. \tag{3.20}$$
- For $m > m_1$, the island allele can be maintained at a stable equilibrium if both potential equilibrium points $p_\pm$ are in the unit interval. We have
$$p_\pm = \frac{2 - 3h \pm \sqrt{(4m(1-2h) + h^2 s)/s}}{2 - 4h}$$
This is always real for $h \le 0.5$. We can exclude this case for the same reasons as in the discrete case. For $h > 0.5$, we need
$$m < m_2 := \frac{h^2 s}{4(2h-1)} \tag{3.21}$$
For $h > 0.5$, $f(p)$ has a single maximum at $\hat{p} = (3hs - 2s)/(4hs - 2s)$. For $h > 2/3$, this maximum is in the unit interval. We have $m_2 \ge m_1$ and $m_1 = m_2$ for $h = 2/3$.
$$\begin{aligned} m &< m_2 \quad \text{for} \quad h \ge \frac{2}{3}\,, \\ m &< m_1 \quad \text{for} \quad h < \frac{2}{3}\,. \end{aligned} \tag{3.22}$$

Note also that $m_{1,2}$ are the leading order approximations for small s of the counterparts in the discrete model. The most significant difference is that the boundary between the two regimes depends only on h in continuous time.

3.4 Two demes with two alleles

We only treat the haploid (diploid multiplicative) model. In discrete time, we get
$$p_1' = (1-m_1)\frac{p_1 w_1(A)}{p_1 w_1(A) + (1-p_1) w_1(a)} + m_1 \frac{p_2 w_2(A)}{p_2 w_2(A) + (1-p_2) w_2(a)} \tag{3.23}$$
$$p_2' = (1-m_2)\frac{p_2 w_2(A)}{p_2 w_2(A) + (1-p_2) w_2(a)} + m_2 \frac{p_1 w_1(A)}{p_1 w_1(A) + (1-p_1) w_1(a)}\,, \tag{3.24}$$

which leads to high-order polynomials for the equilibria. The simpler dynamics in continuous time reads
$$\dot{p}_1 = p_1(w_1(A) - \bar{w}_1) - m_1(p_1 - p_2) = p_1(1-p_1)s_1 - m_1(p_1 - p_2) \tag{3.25}$$
$$\dot{p}_2 = p_2(w_2(A) - \bar{w}_2) - m_2(p_2 - p_1) = p_2(1-p_2)s_2 - m_2(p_2 - p_1)\,, \tag{3.26}$$

where $s_i = w_i(A) - w_i(a)$ is the selection coefficient of the A allele in deme i. Setting $\dot{p}_1 = \dot{p}_2 = 0$, the equation system leads to the condition
$$p_1(1-p_1)\Big((1 - (1-p_1)s_1/m_1)(1 + p_1 s_1/m_1)s_2/m_2 + s_1/m_1\Big) = 0\,.$$

This results in the monomorphic equilibria $\mathbf{p}_0 = (0, 0)$ and $\mathbf{p}_1 = (1, 1)$ as the only boundary equilibria. In addition, there may be an internal equilibrium $\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2)$ with

$$\hat{p}_{1\pm} = \frac{1}{2} - \frac{m_1}{s_1} \pm \sqrt{\frac{1}{4} - \frac{m_1}{s_1}\frac{m_2}{s_2}} \quad ; \quad \hat{p}_{2\mp} = \frac{1}{2} - \frac{m_2}{s_2} \mp \sqrt{\frac{1}{4} - \frac{m_1}{s_1}\frac{m_2}{s_2}}$$

Clearly, an internal equilibrium is only possible if the direction of selection in both demes is opposite, $s_1 s_2 < 0$ since otherwise $\dot{p}_1$ and $\dot{p}_2$ cannot both be zero. From $0 < \hat{p}_1 + \hat{p}_2 < 2$, we further find that

$$\left| \frac{m_1}{s_1} + \frac{m_2}{s_2} \right| < 1 \tag{3.27}$$

is necessary condition. Assuming (without restriction) $s_1 > 0$ and $s_2 < 0$, we have $p_{1-} < 0$ and $p_{2+} > 1$, but $\hat{\mathbf{p}} = (\hat{p}_{1+}, \hat{p}_{2-})$ is in the interior of the frequency space since from (3.27)

$$\frac{m_1}{s_1} - 1 < \left|\frac{m_2}{s_2}\right| < \frac{m_1}{s_1} + 1 \quad ; \quad \left|\frac{m_2}{s_2}\right| - 1 < \frac{m_1}{s_1} < \left|\frac{m_2}{s_2}\right| + 1$$

and thus

$$0 \leq \frac{1}{2} - \frac{m_1}{s_1} + \sqrt{\frac{1}{4} + \frac{m_1}{s_1}\left(\frac{m_1}{s_1} - 1\right)} < \hat{p}_{1+} < \frac{1}{2} - \frac{m_1}{s_1} + \sqrt{\frac{1}{4} + \frac{m_1}{s_1}\left(\frac{m_1}{s_1} + 1\right)} = 1$$

and

$$1 \geq \frac{1}{2} - \frac{m_2}{s_2} - \sqrt{\frac{1}{4} + \left|\frac{m_2}{s_2}\right|\left(\left|\frac{m_2}{s_2}\right| - 1\right)} > \hat{p}_{2-} > \frac{1}{2} - \frac{m_2}{s_2} - \sqrt{\frac{1}{4} + \left|\frac{m_2}{s_2}\right|\left(\left|\frac{m_2}{s_2}\right| + 1\right)} = 0\,.$$

Stability The Jacobian of the dynamical system reads

$$\mathbf{J} = \begin{pmatrix} (1-2p_1)s_1 - m_1 & m_1 \\ m_2 & (1-2p_2)s_2 - m_2 \end{pmatrix}$$

and the maximum eigenvalue follows as

$$\lambda_{\max} = \frac{1}{2}\Big((1-2p_1)s_1 + (1-2p_2)s_2 - m_1 - m_2 \\ + \sqrt{((1-2p_1)s_1 - (1-2p_2)s_2 - m_1 + m_2)^2 + 4m_1 m_2}\Big)\,. \tag{3.28}$$

An equilibrium $\mathbf{p}$ is stable if $\lambda_{\max}(\mathbf{p}) < 0$. For $\mathbf{p} = \mathbf{p}_0$ this requires that $s_1 + s_2 < m_1 + m_2$ and $s_1 m_2 + s_2 m_1 < s_1 s_2$. As expected, this is always the case for $s_1, s_2 < 0$ and never for $s_1, s_2 > 0$ (since $(m_1 + m_2)/(s_1 + s_2) < m_1/s_1 + m_2/s_2$ in the latter case). From symmetry, we conclude that $\mathbf{p}_1$ is stable for $s_1, s_2 > 0$ and unstable for $s_1, s_2 < 0$. In both cases, $m_2 p_1 + m_1 p_2$ is a Lyapunov function (i.e., strictly monotonic under the dynamics) in the interior of the state space. We therefore have global convergence to the respective monomorphic equilibrium. For opposite selection in both demes ($s_1 > 0$ and $s_2 < 0$, say), $\mathbf{p}_0$ is stable iff $m_1/s_1 + m_2/s_2 > 1$ (which implies $m_1 > s_1$ and thus $s_1 + s_2 < m_1 + m_2$). Similarly, $\mathbf{p}_1$ is stable iff $m_1/s_1 + m_2/s_2 < -1$. From condition (3.27) we conclude that both monomorphic equilibria are unstable whenever an internal equilibrium exists. Otherwise, exactly one monomorphic equilibrium is stable. Global stability of the internal equilibrium follows from an analysis of the isolclines of $\dot{p}_1 = 0$ and $\dot{p}_2 = 0$. For $s_1 > 0$ and $s_2 < 0$, both are monotonically increasing functions $p_2(p_1)$ in the interior of the state space. It is then easy to see that their intersection (if it exists) must be a global attractor.

4 Literature

- Durrett R (2008) Probability Models for DNA Sequence Evolution (2nd ed.) Springer.
  A lot of material on advanced stochastic modelling, including selective sweeps.

- Ewens WJ (2004) Mathematical Population Genetics (2nd ed.) Springer.
  Standard reference for the stochastic theory, in particular diffusions.

- Nagylaki T (1992) Introduction to Theoretical Population Genetics. Springer.
  Comprehensive account in particular of the deterministic theory. Chapter 6 on migration-selection models.

- Wakeley J (2008) Coalescent Theory: An Introduction. Roberts & Company Publishers, Greenwood Village, Colorado.
  Standard introductory textbook on coalescent theory.